Psychological Methods
2003, Vol. 8, No. 4, 448–467

1082-989X/03/$12.00 DOI: 10.1037/1082-989X.8.4.448

Effect-Size Indices for Dichotomized Outcomes in Meta-Analysis

Julio Sánchez-Meca and
Fulgencio Marín-Martínez
University of Murcia

Salvador Chacón-Moscoso
University of Seville

It is very common to find meta-analyses in which some of the studies compare 2 groups on continuous dependent variables and others compare groups on dichotomized variables. Integrating all of them in a meta-analysis requires an effect-size index in the same metric that can be applied to both types of outcomes. In this article, the performance in terms of bias and sampling variance of 7 different effect-size indices for estimating the population standardized mean difference from a 2 × 2 table is examined by Monte Carlo simulation, assuming normal and nonnormal distributions. The results show good performance for 2 indices, one based on the probit transformation and the other based on the logistic distribution.

In the last 20 years, meta-analysis has become a very popular and useful research methodology to integrate the results of a set of empirical studies about a given topic. To carry out a meta-analysis an effect-size index has to be selected to translate the results of every study into a common metric.

When the focus of a study is to compare the performance of two groups (e.g., treated vs. control, male vs. female, trained vs. nontrained) on a continuous dependent variable, the effect-size index most usually applied is the *standardized mean difference, d,* defined as the difference between the means of the two groups divided by a within-group standard deviation estimate. Substantial meta-analytic literature has been devoted to showing the properties of this index, its sampling variance, how to obtain confidence intervals on the weighted mean effect size obtained from it, its statistical significance, homogeneity tests, and how to search for variables that moderate it (Cooper, 1998; Cooper & Hedges, 1994; Hedges & Olkin, 1985; Lipsey & Wilson, 2001; Rosenthal, 1991).

However, in social and behavioral research it is also very common to find studies using dichotomous outcome measures or continuous variables that have been dichotomized. Here we focus on *dichotomized variables,* that is to say, variables that represent constructs continuous in nature, although being measured as a dichotomy. This definition includes studies in which the researchers apply a cutpoint, Y_c, on a quantitative variable, as well as those in which they directly measure the dependent variable as a dichotomy although there is an underlying continuous construct. For many outcome measures, this is a reasonable assumption. For example, studies about the effectiveness of the treatment of tobacco addiction can measure the results as the number of cigarettes smoked in a day or as a dichotomy (tobacco abstinence vs. nonabstinence). In the field of delinquency treatment, studies can record recidivism into crime as a dichotomy (recidivist vs. nonrecidivist) or, for example, as the number of police contacts after prison release. In education, the performance of students on an exam can be measured continuously as the number of points scored or dichotomously as passing versus failing the exam.

Julio Sánchez-Meca and Fulgencio Marín-Martínez, Department of Basic Psychology and Methodology, University of Murcia, Murcia, Spain; Salvador Chacón-Moscoso, Department of Psychology, University of Seville, Seville, Spain.

This article was supported by a grant from the Ministerio de Ciencia y Tecnología and by funds from the Fondo Europeo de Desarrollo Regional (FEDER; Project Number BSO2001-0491). We gratefully acknowledge William R. Shadish and Mark W. Lipsey for their helpful suggestions on a first draft of this article. We also thank the referees for comments that greatly improved the manuscript.

Correspondence concerning this article should be addressed to Julio Sánchez-Meca, Department of Basic Psychology and Methodology, Faculty of Psychology, Campus of Espinardo, University of Murcia, P.O. Box 4021, 30100 Murcia, Spain. E-mail: jsmeca@um.es

In these cases, the study results can be summarized as a 2 × 2 contingency table, where two groups are crossed with the two outcomes, giving four possible cell frequencies, as shown in Table 1. With n_E and n_C being the sample sizes, O_{1E} and O_{1C} being the success frequencies, and $p_E = O_{1E}/n_E$ and $p_C = O_{1C}/n_C$ being the success proportions in the experimental and control groups, respectively, different effect-size indices have been proposed to represent the effect magnitude (Fleiss, 1981, 1994; Laird & Mosteller, 1990; Lipsey & Wilson, 2001; Rosenthal, 1994, 2000; Shadish & Haddock, 1994). The first purpose of the present article is to explore the properties of these indices.

Three effect-size indices have been applied often: (a) the *risk difference,* $p_E - p_C$, the raw difference between the two success (or failure) proportions; (b) the *risk ratio,* p_E/p_C, the ratio between the two proportions, and (c) the *odds ratio,* $p_E(1 - p_C)/p_C(1 - p_E)$, the relative odds that one will be more successful than the other. Of the three indices, the odds ratio is the best one for most situations because of its good statistical properties (Fleiss, 1994; Haddock, Rindskopf, & Shadish, 1998), although risk difference and risk ratio can also be good alternatives under certain conditions (Deeks & Altman, 2001; Hasselblad, Mosteller, et al., 1995; Sánchez-Meca & Marín-Martínez, 2000, 2001). In particular, these three indices have been applied in meta-analyses in the health sciences, because in this field it is very common to find research issues in which the outcome is always measured as a dichotomous (or dichotomized) variable.

More commonly, some of the studies in a meta-analysis present results comparing the performance of two groups on continuous outcome variables, other studies present them on dichotomized variables, and some studies include both continuous and dichotomized variables. In these cases, if some effect-size indices are computed as *d,* and others as an odds ratio, risk ratio, or risk difference, some means of converting all these diverse indices to a common effect size is necessary to integrate all the results into a single average effect size. Consequently, a second purpose of the present article is to evaluate different methods for converting different effect-size indices into the *d* metric.[1]

These cases are often handled in one of the following ways. A dichotomous version of the standardized mean difference that has been commonly applied consists of calculating the difference between success (or failure) proportions in experimental and control groups and dividing it by an estimate of within-group standard deviation (Fleiss, 1981); this index could be named the *standardized proportion difference,* d_p, and in many cases it underestimates the population standardized mean difference (Fleiss, 1994; Haddock et al., 1998).

Another strategy consists of computing the phi coefficient, ϕ, from each one of the 2 × 2 tables and sometimes also transforming it into a standardized mean difference, d_ϕ, by means of the typical *r* to *d* translation formulas (e.g., Hedges & Olkin, 1985; Rosenthal, 1991). In this case, phi coefficients also underestimate the population correlation coefficient and, therefore, d_ϕ indices would also underestimate the meta-analytic results (Fleiss, 1994; Haddock et al., 1998). Conversely, some meta-analysts transform ev-

Table 1
Contingency 2 × 2 Table for Two Groups and a Dichotomized Outcome

Outcome	Group		Total
	Experimental	Control	
Success ($Y_i \geq Y_c$)	O_{1E}	O_{1C}	m_1
Failure ($Y_i < Y_c$)	O_{2E}	O_{2C}	m_2
Total	n_E	n_C	N

Note. Y_i = the continuous outcome variable; Y_c = the cutpoint applied for dichotomizing the dependent variable, *Y;* O_{1E} and O_{1C} = the success frequencies of the experimental and control groups, respectively; O_{2E} and O_{2C} = the failure frequencies of the experimental and control groups, respectively; $m_1 = O_{1E} + O_{1C}$; $m_2 = O_{2E} + O_{2C}$; n_E and n_C = the sample sizes for experimental and control groups, respectively. $N = n_E + n_C$.

[1] Whitehead, Bailey, and Elbourne (1999) proposed another strategy consisting of estimating the log odds ratio in each study with continuous measures assuming normal (or log-normal) distributions. The method requires a cutpoint, Y_c, for obtaining the standard normal (or log-normal) distribution function for experimental and control groups, p_E and p_C, in every study. Then, the log odds ratio estimates so obtained can be quantitatively integrated with the log odds ratios obtained from the dichotomous outcomes. This strategy is especially useful when all of the studies with continuous outcomes included in the meta-analysis have used the same scale, allowing the same cutpoint to be applied in all of the studies. However, when the different studies have used different measures and scales, it would be difficult (and arbitrary) to define the cutpoints needed for estimating the log odds ratios. Another problem in this strategy is the loss of information produced in the process of dichotomizing variables (Dominici & Parmigiani, 2000). Therefore, provided that meta-analyses in educational and behavioral sciences routinely include different measures and scales of the same construct, this strategy is not practical.

ery standardized mean difference, *d,* obtained from studies with continuous outcome variables into the Pearson correlation coefficient, *r,* and then integrate them with phi coefficients obtained from studies with 2 × 2 tables. This will also underestimate the population correlation coefficient.

Only recently have the problems of the phi coefficient and standardized-proportion-difference statistics applied to 2 × 2 tables been discussed in the social and behavioral sciences (Fleiss, 1994; Haddock et al., 1998; Lipsey & Wilson, 2001). Currently, several alternative strategies better than those involving d_p and d_ϕ indices can be applied to transform different effect-size indices into the *d* metric. One of the strategies is the correction of the attenuation in the d_p and d_ϕ indices that arises because of the dichotomization of the underlying continuous variable (Becker & Thorndike, 1988; Hunter & Schmidt, 1990; Lipsey & Wilson, 2001). Other strategies are based on the assumption of the logistic distribution or on the arcsine transformation. In this article we compare the performance of seven translation formulas that put a variety of indices on the *d* scale.

A third question addressed in this article is that of the sampling variance of the effect-size indices. In meta-analysis, such sampling variances are important because the meta-analytic statistical models usually weight each effect size by its inverse variance. When the dependent variable is dichotomized, a loss of information is produced that affects the accuracy of the effect size. The formulas derived from statistical theory for estimating the sampling variance of each effect-size index have to reflect the cost of dichotomization in terms of accuracy. Therefore, it is expected that the sampling variances of the transformed indices will be larger than that of the standardized mean difference, *d.*

In summary, then, this article uses Monte Carlo simulation to examine the performance of seven different strategies for obtaining a *d* index when the dependent variable has been dichotomized. On the basis of past research, the normal distribution is the most usual assumption for data in the empirical studies; so we compared the bias and the sampling variance of the different effect-size indices assuming that the two populations are normally distributed. To check the robustness of these effect estimators, we have also included several conditions representing nonnormal distributions. Several factors in the simulation were manipulated: the population standardized mean difference, δ, the value of the cutpoint, Y_c, to dichotomize the distributions, the sample size, the imbalance between sample sizes, and the relationship between sample sizes and success proportions. It was expected that d_p and d_ϕ would underestimate the population standardized mean difference, that the remaining indices would offer a better performance, and that indices based on normal distributions would present the best results when normality is assumed. In any case, the influence of varying all of these factors on the performance of these effect-size indices is a question that has not been yet studied.

Effect-Size Indices for Summarizing 2 × 2 Tables

We assume that the population contains two continuous distributions (those of experimental and control groups) with μ_E and μ_C as experimental and control population means, respectively, with σ being the common population standard deviation. Thus, the parametric effect size between experimental and control groups is defined as the standardized mean difference, δ, and is computed as

$$\delta = \frac{\mu_E - \mu_C}{\sigma} \tag{1}$$

(Hedges & Olkin, 1985, Equation 2, p. 76).

We assume that the continuous variable is normally distributed for experimental and control groups [$Y_{iE} \sim N(\mu_E, \sigma^2)$; $Y_{iC} \sim N(\mu_C, \sigma^2)$]. When in a single study the outcome variable has been measured continuously, the parametric effect size, δ, can be estimated by means of the sample standardized mean difference, *g,* computed by

$$g = \frac{\bar{y}_E - \bar{y}_C}{S} \tag{2}$$

(Hedges & Olkin, 1985, Equation 3, p. 78), with $\bar{y}_E$ and $\bar{y}_C$ being the sample means of the experimental and control groups, and *S* being a pooled estimate of the within-group standard deviation, given by

$$S = \sqrt{\frac{(n_E - 1)S_E^2 + (n_C - 1)S_C^2}{n_E + n_C - 2}} \tag{3}$$

(Hedges & Olkin, 1985, p. 79), with S_E^2 and S_C^2 being the sample variances of the experimental and control groups, respectively. To correct the positive bias of

the standardized mean difference for small sample sizes, the correction factor proposed in Hedges and Olkin (1985, Equation 10, p. 81) would be applied to the g index to obtain an unbiased estimate, d, of δ:

$$d = c(m)g, \tag{4}$$

where $c(m)$ is the correction factor and is obtained by

$$c(m) = 1 - \frac{3}{4m - 1}, \tag{5}$$

with $m = n_E + n_C - 2$. The sampling variance of the d index is estimated by

$$S_d^2 = \frac{n_E + n_C}{n_E n_C} + \frac{d^2}{2(n_E + n_C)} \tag{6}$$

(Hedges & Olkin, Equation 15, p. 86).

In the case of studies with 2 × 2 tables, it is supposed that the continuous outcome variable has been dichotomized by applying some cutpoint, Y_c, to the two original continuous populations, to classify the subjects of the two populations into success or failure categories. Two extensions of the standardized mean difference d have been applied to studies with such 2 × 2 tables: the standardized proportion differences, d_p, obtained from success (or failure) proportions, and the phi coefficient. The former is defined as the difference between success (or failure) proportions in experimental and control groups (p_E and p_C, respectively), divided by an estimate of the within-group standard deviation, S'. This index, d_p, is computed as

$$d_p = \frac{p_E - p_C}{S'} \tag{7}$$

(Johnson, 1989, p. 150; Johnson & Eagly, 2000, p. 511), where S' is given as

$$S' = \sqrt{\frac{(n_E - 1)p_E(1 - p_E) + (n_C - 1)p_C(1 - p_C)}{n_E + n_C - 2}} \tag{8}$$

As Haddock et al. (1998) stated, and as shown in our simulation study, d_p underestimates the effect in the population, δ, whenever the marginals are not proportional. The usual estimate of its sampling variance, $S_{d_p}^2$, is obtained by

$$S_{d_p}^2 = \frac{n_E + n_C}{n_E n_C} + \frac{d_p^2}{2(n_E + n_C)}. \tag{9}$$

The phi coefficient ϕ is the other effect-size index usually applied to 2 × 2 tables. It is computed as

$$\phi = \frac{O_{1E}O_{2C} - O_{2E}O_{1C}}{\sqrt{n_E n_C m_1 m_2}} \tag{10}$$

(e.g., Fleiss, 1994, Equation 17-11, p. 249), with all of the terms in the equation defined in Table 1. Phi can be translated to the metric of the standardized mean difference, d_ϕ, by

$$d_\phi = \frac{\phi}{\sqrt{1 - \phi^2}} \sqrt{\frac{df(n_E + n_C)}{n_E n_C}} \tag{11}$$

(Rosenthal, 1994, Equation 16-29, p. 239), with $df = n_E + n_C - 2$.

The phi coefficient and, consequently, d_ϕ underestimate the parametric effect size (Haddock et al., 1998). The sampling variance of d_ϕ can be approached by applying the delta method (Stuart & Ord, 1994, p. 350) to Equation 11:

$$S_{d_\phi}^2 = \frac{n_E + n_C}{n_E n_C (1 - \phi^2)^2}. \tag{12}$$

Cohen (1988, p. 181) proposed the arcsine translation of the success proportion in a 2 × 2 table, p_E and p_C, for obtaining an effect size in the d metric:

$$d_{asin} = 2\text{arcsine}\sqrt{p_E} - 2\text{arcsine}\sqrt{p_C}. \tag{13}$$

Following Lipsey and Wilson (2001, p. 56), we expected that d_{asin} would underestimate the population effect size, unless the population distributions were very skew. The sampling variance of d_{asin} is approximated by

$$S_{d_{asin}}^2 = \frac{1}{n_E} + \frac{1}{n_C}. \tag{14}$$

(Rosenthal, 1994, p. 238).

A fourth index is based on the odds ratio. Assuming logistic distributions and homogeneous variances, Hasselblad and Hedges (1995, Equation 5, p. 170; see also Chinn, 2000) proposed transforming the log odds ratio into d by

$$d_{HH} = L_{OR}\frac{\sqrt{3}}{\pi}, \tag{15}$$

where $\pi = 3.14159$, L_{OR} is the natural logarithm of the odds ratio (OR); the odds ratio is easily obtained from the 2 × 2 table by

$$OR = \frac{p_E (1 - p_C)}{p_C (1 - p_E)} \tag{16}$$

(e.g., Shadish & Haddock, 1994, Equation 18-11, p. 269), with the precaution of adding 0.5 to all cell frequencies when any of them is 0. Under these assumptions, the log odds ratio is just the constant $\pi/\sqrt{3} = 1.81$ multiplied by the standardized mean difference, *d*. Therefore, the d_{HH} index is obtained dividing the log odds ratio by the constant 1.81, which is also the standard deviation of the logistic distribution. Applying the d_{HH} index under the normal distribution assumption probably will slightly underestimate the population standardized mean difference, δ.

If each of the two continuous populations follows a logistic distribution with equal variances, the d_{HH} index is independent of the cutpoint Y_c and is normally distributed, provided that O_{1E}, O_{1C}, O_{2E}, and O_{2C} are not too small. Applying the delta method we can approximate the sampling variance of d_{HH} by

$$S^2_{d_{HH}} = \frac{3}{\pi^2}\left[\frac{1}{O_{1E}} + \frac{1}{O_{2E}} + \frac{1}{O_{1C}} + \frac{1}{O_{2C}}\right] \tag{17}$$

(Hasselblad & Hedges, 1995).

An effect size similar to d_{HH} was proposed by Cox (1970) and cited in Haddock et al. (1998). It consists of dividing the log odds ratio by the constant 1.65:

$$d_{Cox} = L_{OR}/1.65, \tag{18}$$

and its sampling variance is estimated as

$$S^2_{d_{Cox}} = 0.367\left[\frac{1}{O_{1E}} + \frac{1}{O_{2E}} + \frac{1}{O_{1C}} + \frac{1}{O_{2C}}\right]. \tag{19}$$

However, most primary studies assume a normal distribution in the underlying populations, and under these conditions, the performance of the preceding effect-size indices based on the logit transformation can differ. Two effect-size indices based on the normal distribution assumption are the *probit transformation* and the *biserial-phi coefficient.* Glass, McGaw, and Smith (1981) proposed the probit transformation to obtain an effect-size index in the *d* metric. Let p_E and p_C be the success proportions in experimental and control groups, respectively, obtained from the 2 × 2 table in a given study. The probit transformation, d_{Probit}, is obtained by

$$d_{Probit} = (z_E - z_C) \tag{20}$$

(Glass et al., 1981, p. 138), z_E and z_C being the inverse of the standard normal distribution function for p_E and p_C, respectively [$z_E = \Phi^{-1}(p_E)$; $z_C = \Phi^{-1}(p_C)$]. Assuming normal distributions, d_{Probit} will be an unbiased estimator of the population standardized mean difference. The sampling variance of d_{Probit} is estimated as

$$S^2_{d_{Probit}} = \left[\frac{2\pi p_E(1 - p_E)e^{z_E^2}}{n_E} + \frac{2\pi p_C(1 - p_C)e^{z_C^2}}{n_C}\right] \tag{21}$$

(Rosenthal, 1994, p. 238).

Another option that assumes normal distributions to summarize the results of a 2 × 2 table consists of calculating the biserial-phi correlation coefficient, ϕ_{bis}, and translating it into the *d* index by a typical *r* to *d* translation formula (e.g., Hedges & Olkin, 1985; Rosenthal, 1991). This correlation coefficient was proposed by Thorndike (1949, 1982; see also Becker & Thorndike, 1988) in the field of psychometrics. The biserial-phi correlation is an unbiased estimate of the point-biserial correlation when the continuous variable has been dichotomized. It can be also defined as a correction of the underestimation produced by the phi coefficient, this correction being the same as that proposed in Hunter and Schmidt (1990; see also Lipsey & Wilson, 2001, p. 111). The estimator is created by multiplying the phi coefficient by the same multiplier used in creating the biserial. Therefore, the biserial-phi coefficient can be obtained from the 2 × 2 table by

$$\phi_{bis} = \frac{\sqrt{p'q'}}{y'}\phi, \tag{22}$$

Becker & Thorndike, 1988, p. 525), where p' is the global success proportion in the 2 × 2 table $p' = (O_{1E} + O_{1C})/N$; y' is the probability density function of the standard normal distribution corresponding to p'; and $q' = 1 - p'$. Following the same strategy used with the phi coefficient, ϕ_{bis} is translated to a *d* index, d_{bis}, by means of

$$d_{bis} = \frac{\phi_{bis}}{\sqrt{1 - \phi^2_{bis}}}\sqrt{\frac{df(n_E + n_C)}{n_E n_C}} \tag{23}$$

(Rosenthal, 1994, Equation 16–29, p. 239), with $df = n_E + n_C - 2$. It is expected that d_{bis} will be an unbiased estimator of the population standardized mean difference. The sampling variance of d_{bis} is estimated by means of the delta method as

$$S^2_{d_{bis}} = \frac{p'q'(1-\phi^2)(n_E+n_C)}{y'^2 n_E n_C (1-\phi^2_{bis})^3}. \quad (24)$$

Because d, d_p, d_ϕ, d_{asin}, d_{HH}, d_{Cox}, d_{Probit}, and d_{bis} are in the same metric, their statistical properties can be compared. Although we already know that d_p and d_ϕ underestimate the population effect size δ, neither the magnitude of the bias, nor how different factors can affect d_p and d_ϕ, has been explored. On the other hand, studying the performance of d_{HH} when the logistic distribution assumption is not met enables us to determine the appropriateness of this index under other distributions. Moreover, under normal distributions it is expected that d_{Probit} and d_{bis} will offer the best results. Note that the results of our simulation study are conditioned by the normal distribution assumption. We have assumed such a condition because most primary studies make this assumption. In fact, although the normality assumption is often not tested, researchers routinely apply parametric statistical tests based on normal distributions. However, when the meta-analyst considers that logistic or other distributions are more realistic, then the results of our simulations should be interpreted very cautiously. To explore the performance of these effect indices under nonnormal distributions, we have added a few conditions representing skewed distributions.

An Example

To clarify the computational formulas of the different effect-size indices proposed and their sampling variances, we have developed a numerical example by randomly generating two samples of $n_E = n_C = 20$ from two normal distributions [$Y_{iE} \sim N(42, 16)$; $Y_{iC} \sim N(40, 16)$] with parametric effect size, $\delta = (42 - 40)/4 = 0.50$. Applying the cutpoint $Y_c = 41$ for classifying, into success versus failure, the scores of the two groups, we obtained the 2×2 table presented in Table 2. Tables 3 and 4 present the calculations for the seven effect-size indices and their sampling variances, respectively.

As expected, the three indices that showed the

Table 2
Data of the Example Obtained by Random Sampling From Two Normal Distributions

Case	Group	
	Experimental	Control
1	36.318	33.992
2	39.842	33.664
3	37.813	40.518
4	36.210	40.451
5	53.967	36.719
6	46.811	36.042
7	41.626	40.681
8	47.580	42.241
9	38.416	49.573
10	44.624	40.998
11	45.985	34.070
12	40.441	40.610
13	44.281	39.058
14	36.705	42.612
15	40.178	39.845
16	41.058	42.247
17	42.951	42.126
18	40.368	37.482
19	42.643	35.677
20	38.113	43.145
$\bar{y}_j$	41.796	39.587
S_j	4.468	3.877

Calculations of d index and its sampling variance

$$S = \sqrt{\frac{(n_E-1)S^2_E + (n_C-1)S^2_C}{n_E+n_C-2}}$$

$$= \sqrt{\frac{(20-1)4.468^2 + (20-1)3.877^2}{20+20-2}} = 4.183$$

$$g = \frac{\bar{y}_E - \bar{y}_C}{S} = \frac{41.796 - 39.587}{4.183} = 0.528$$

$$c(m) = 1 - \frac{3}{4m-1} = 1 - \frac{3}{4(20+20-2)-1} = 0.9801$$

$$d = c(m)g = (0.9801)(0.528) = 0.517$$

$$S^2_d = \frac{n_E+n_C}{n_E n_C} + \frac{d^2}{2(n_E+n_C)} = \frac{20+20}{(20)(20)} + \frac{0.517^2}{2(20+20)} = 0.1033$$

Process of dichotomization

Outcome	Group		Total
	Experimental	Control	
$Y_c \geq 41$	10	6	16
$Y_c < 41$	10	14	24
Total	20	20	40

Table 3
Calculations of the Different Effect-Size Indices of the Example in Table 2

d_p index

$$p_E = O_{1E}/n_E = 10/20 = 0.5$$

$$p_C = O_{1C}/n_C = 6/20 = 0.3$$

$$S' = \sqrt{\frac{(n_E - 1)\, p_E\,(1 - p_E) + (n_C - 1) p_C\,(1 - p_C)}{n_E + n_C - 2}} = \sqrt{\frac{(19)(0.5)(0.5) + (19)(0.3)(0.7)}{20 + 20 - 2}} = 0.479$$

$$d_p = \frac{p_E - p_C}{S'} = \frac{0.5 - 0.3}{0.479} = 0.417$$

d_ϕ index

$$\phi = \frac{O_{1E}O_{2C} - O_{2E}O_{1C}}{\sqrt{n_E n_C m_1 m_2}} = \frac{(10)(14) - (6)(10)}{\sqrt{(20)(20)(16)(24)}} = 0.204$$

$$d_\phi = \frac{\phi}{\sqrt{1 - \phi^2}} \sqrt{\frac{df(n_E + n_C)}{n_E n_C}} = \frac{0.204}{\sqrt{1 - 0.204^2}} \sqrt{\frac{38(20 + 20)}{(20)(20)}} = 0.406$$

d_{asin} index

$$d_{asin} = 2\text{arcsine}\sqrt{p_E} - 2\text{arcsine}\sqrt{p_C} = 2\text{arcsine}\sqrt{0.5} - 2\text{arcsine}\sqrt{0.3} = 0.411$$

d_{HH} index

$$OR = \frac{p_E\,(1 - p_C)}{p_C\,(1 - p_E)} = \frac{(0.5)(1 - 0.3)}{(0.3)(1 - 0.5)} = 2.333$$

$$L_{OR} = Log_e\,(OR) = Log_e\,(2.333) = 0.847$$

$$d_{HH} = L_{OR}\frac{\sqrt{3}}{\pi} = (0.847)\frac{\sqrt{3}}{3.14159} = 0.467$$

$d_{Cox\ index}$

$$d_{Cox} = L_{OR}/1.65 = 0.847/1.65 = 0.513$$

d_{Probit} index

$$z_E = \Phi^{-1}\,(p_E) = \Phi^{-1}\,(0.5) = 0.0$$

$$z_C = \Phi^{-1}\,(p_C) = \Phi^{-1}\,(0.3) = -0.524$$

$$d_{Probit} = (z_E - z_C) = 0.0 - (-0.524) = 0.524$$

d_{bis} index

$$p' = (O_{1E} + O_{1C})/N = (10 + 6)/40 = 0.4$$

$$q' = 1 - p' = 1 - 0.4 = 0.6$$

$$y' = 0.3864$$

$$\phi_{bis} = \frac{\sqrt{p'q'}}{y'}\phi = \frac{\sqrt{(0.4)(0.6)}}{0.3864}(0.204) = 0.259$$

$$d_{bis} = \frac{\phi_{bis}}{\sqrt{1 - \phi_{bis}^2}} \sqrt{\frac{df\,(n_E + n_C)}{n_E n_C}} = \frac{0.259}{\sqrt{1 - 0.259^2}} \sqrt{\frac{38(20 + 20)}{(20)(20)}} = 0.523$$

Table 4
Calculations of the Sampling Variance of the Effect-Size Indices of the Example in Table 2

d_p index

$$S^2_{d_p} = \frac{n_E + n_C}{n_E n_C} + \frac{d_p^2}{2(n_E + n_C)} = \frac{20 + 20}{(20)(20)} + \frac{0.417^2}{2(20 + 20)} = 0.1022$$

d_ϕ index

$$S^2_{d_\phi} = \frac{n_E + n_C}{n_E n_C (1 - \phi^2)^2} = \frac{20 + 20}{(20)(20)(1 - 0.204^2)^2} = 0.1089$$

d_{asin} index

$$S^2_{d_{asin}} = \frac{1}{n_E} + \frac{1}{n_C} = \frac{1}{20} + \frac{1}{20} = 0.10$$

d_{HH} index

$$S^2_{d_{HH}} = \frac{3}{\pi^2}\left[\frac{1}{O_{1E}} + \frac{1}{O_{2E}} + \frac{1}{O_{1C}} + \frac{1}{O_{2C}}\right] = \frac{3}{3.14159^2}\left[\frac{1}{10} + \frac{1}{10} + \frac{1}{6} + \frac{1}{14}\right] = 0.1332$$

d_{Cox} index

$$S^2_{d_{Cox}} = 0.367\left[\frac{1}{O_{1E}} + \frac{1}{O_{2E}} + \frac{1}{O_{1C}} + \frac{1}{O_{2C}}\right] = 0.367\left[\frac{1}{10} + \frac{1}{10} + \frac{1}{6} + \frac{1}{14}\right] = 0.1608$$

d_{Probit} index

$$S^2_{d_{Probit}} = \left[\frac{2\pi p_E (1 - p_E) e^{z_E^2}}{n_E} + \frac{2\pi p_C (1 - p_C) e^{z_C^2}}{n_C}\right]$$
$$= \left[\frac{(2)(3.14159)(0.5)(0.5)e^{0^2}}{20} + \frac{(2)(3.14159)(0.3)(0.7)e^{-0.524^2}}{20}\right] = 0.1653$$

d_{bis} index

$$S^2_{d_{bis}} = \frac{p'q'(1 - \phi^2)(n_E + n_C)}{y'^2 n_E n_C (1 - \phi^2_{bis})^3} = \frac{(0.4)(0.6)(1 - 0.204^2)(20 + 20)}{0.3864^2 (20)(20)(1 - 0.259^2)^3} = 0.1898$$

highest underestimations of δ were $d_p = 0.417$, $d_\phi = 0.406$, and $d_{asin} = 0.411$, followed by $d_{HH} = 0.467$ (Table 3). The estimates closest to δ were $d_{Cox} = 0.513$, $d_{Probit} = 0.524$, and $d_{bis} = 0.523$. With respect to the sampling variances (Table 4), those of d_p ($S^2_{d_p} = 0.1022$), d_ϕ ($S^2_{d_\phi} = 0.1089$), and d_{asin} ($S^2_{d_{asin}} = 0.10$) were very close to that of the d index ($S^2_d = 0.1033$) and were smaller than those of d_{HH} ($S^2_{d_{HH}} = 0.1332$), d_{Cox} ($S^2_{d_{Cox}} = 0.1608$), d_{Probit} ($S^2_{d_{Probit}} = 0.1653$), and d_{bis} ($S^2_{d_{bis}} = 0.1898$). Taking in the example the sampling variance of d as the point of reference, the loss of efficiency due to the dichotomization in d_{HH}, d_{Cox}, d_{Probit}, and d_{bis} indices is not trivial: 28.9%, 55.7%, 60.0%, and 83.7%, respectively.

Method

The simulation study was programmed in Gauss (Aptech Systems, 1992). Two normally distributed populations with homogeneous variances were defined, $N(\mu_E, \sigma^2)$ and $N(\mu_C, \sigma^2)$, where μ_E and μ_C are the experimental and control population means, respectively, and σ is the common standard deviation. The population standardized mean difference, δ, was

defined in Equation 1. We assumed $\sigma^2 = 1$, $\mu_C = 0$, and consequently, $\mu_E = \delta$. Furthermore, one cutpoint, Y_c, dichotomized the continuous variable into two levels. Pairs of independent random samples of sizes n_E and n_C were generated from these populations.

Each pair of generated samples simulated the data in a primary research study, which we expressed in two metrics: the quantitative scores and the data dichotomized into two levels, over and under the cutpoint, forming a 2 × 2 table. When any of the cell frequencies in the 2 × 2 table was zero, 0.5 was added to all of them to avoid problems in the computation of the effect indices. For the quantitative scores the d index and its sampling variance were computed (Equations 4 and 6). For the dichotomized data the d_p, d_ϕ, d_{asin}, d_{HH}, d_{Cox}, d_{Probit}, and d_{bis} indices were computed (Equations 7, 11, 13, 15, 18, 20, and 23, respectively) along with their sampling variances (Equations 9, 12, 14, 17, 19, 21, and 24, respectively).

The following factors were manipulated in the simulations: (a) the total sample size of each study, $N = n_E + n_C$, with values 48, 60, and 100; (b) the ratio between sample sizes of the two groups in each study, with three conditions $n_E = n_C$, $n_E = 2n_C$, and $n_E = 4n_C$; (c) in cases of unequal sample size, whether the experimental or control groups had the largest sample size; (d) following Cohen (1988), the value of the population standardized mean difference, with values of $\delta = 0.2$, 0.5, and 0.8; and (e) the value of the cutpoint. The cutpoint was manipulated in the following way: for $\delta = 0.2$, $Y_c = 0.1$; for $\delta = 0.5$, $Y_c = 0.1$, 0.25, and 0.4; and for $\delta = 0.8$, $Y_c = 0.1$, 0.4, and 0.7. To simplify the presentation of the results, we completely crossed all of the manipulated factors for only $\delta = 0.5$, as it can be considered the effect of a medium magnitude following Cohen (1988). So, only 71 of the 105 possible combinations among the different factors manipulated were reported (see Table 5). In particular, several combinations of ratios between sample sizes, the cutpoint, and the relation between sample size and the experimental and control means were excluded. Additional analyses across the full set of 105 conditions showed results with the same trends found to the 71 conditions reported here.

For each one of the 71 conditions, 10,000 replications were generated. The d, d_p, d_ϕ, d_{asin}, d_{HH}, d_{Cox}, d_{bis}, and d_{Probit} indices were computed in each of these replications. The bias of each of the eight indices was assessed as the difference between the mean of the 10,000 empirical values of each index and the population standardized mean difference δ. On the other hand, the variability of the indices was assessed by the mean squared difference of each of the eight indices with respect to δ, across the 10,000 replications of the same condition. Finally, the formulas for estimating the sampling variances of the eight indices were computed in each replication (Equations 6, 9, 12, 14, 17, 19, 21, and 24), and their values averaged over the 10,000 replicates of the same condition.

To begin an exploration of the performance of the effect indices under departures from the normality assumption, we included nine additional conditions in

Table 5
Conditions Manipulated in the Simulation Study Under Normal Distributions

δ	N	n_E	n_C	Y_c	δ	N	n_E	n_C	Y_c
0.2	48	24	24	.1	0.5	48	16	32	.4
0.2	48	32	16	.1	0.5	60	30	30	.4
0.2	60	30	30	.1	0.5	60	40	20	.4
0.2	60	40	20	.1	0.5	60	20	40	.4
0.2	60	48	12	.1	0.5	60	48	12	.4
0.2	100	50	50	.1	0.5	60	12	48	.4
0.2	100	66	34	.1	0.5	100	50	50	.4
0.2	100	80	20	.1	0.5	100	66	34	.4
0.5	48	24	24	.1	0.5	100	34	66	.4
0.5	48	32	16	.1	0.5	100	80	20	.4
0.5	48	16	32	.1	0.5	100	20	80	.4
0.5	60	30	30	.1	0.8	48	24	24	.1
0.5	60	40	20	.1	0.8	48	32	16	.1
0.5	60	20	40	.1	0.8	60	30	30	.1
0.5	60	48	12	.1	0.8	60	40	20	.1
0.5	60	12	48	.1	0.8	60	48	12	.1
0.5	100	50	50	.1	0.8	100	50	50	.1
0.5	100	66	34	.1	0.8	100	66	34	.1
0.5	100	34	66	.1	0.8	100	80	20	.1
0.5	100	80	20	.1	0.8	48	24	24	.4
0.5	100	20	80	.1	0.8	48	32	16	.4
0.5	48	24	24	.25	0.8	60	30	30	.4
0.5	48	32	16	.25	0.8	60	40	20	.4
0.5	48	16	32	.25	0.8	60	48	12	.4
0.5	60	30	30	.25	0.8	100	50	50	.4
0.5	60	40	20	.25	0.8	100	66	34	.4
0.5	60	20	40	.25	0.8	100	80	20	.4
0.5	60	48	12	.25	0.8	48	24	24	.7
0.5	60	12	48	.25	0.8	48	32	16	.7
0.5	100	50	50	.25	0.8	60	30	30	.7
0.5	100	66	34	.25	0.8	60	40	20	.7
0.5	100	34	66	.25	0.8	60	48	12	.7
0.5	100	80	20	.25	0.8	100	50	50	.7
0.5	100	20	80	.25	0.8	100	66	34	.7
0.5	48	24	24	.4	0.8	100	80	20	.7
0.5	48	32	16	.4					

which the shape of the continuous variable distributions was manipulated. Three levels of population nonnormality were considered: skewness = 0.5 and kurtosis = 0, skewness = 0.75 and kurtosis = 0, and skewness = 1.75 and kurtosis = 3.75. The distribution shapes were identical for the two populations simulated. The total sample size was fixed with $N = 60$, with $n_E = n_C = 30$, for all these conditions; the population standardized mean difference was manipulated with values $\delta = 0.2$, 0.5, and 0.8; and the cutpoint was manipulated with these values: for $\delta = 0.2$, $Y_c = 0.1$; for $\delta = 0.5$, $Y_c = 0.25$; and for $\delta = 0.8$, $Y_c = 0.4$.

Using the Fleishman (1978) power transformation, $X = a + bZ + cZ^2 + dZ^3$, we transformed two standard normal distributions, $Z \sim N(0, 1)$, to reflect the target distribution shapes. For the skewed–mesokurtic distributions (skewness = 0.5 and kurtosis = 0, skewness = 0.75 and kurtosis = 0), the constants were $a = -0.093$, $b = 1.040$, $c = 0.093$, and $d = -0.016$; and $a = -0.174$, $b = 1.114$, $c = 0.174$, and $d = -0.503$, respectively. For the skewed–leptokurtic distribution (skewness = 1.75 and kurtosis = 3.75), the constants were $a = -0.399$, $b = 0.930$, $c = 0.399$, and $d = -0.036$. To generate data with $\sigma^2 = 1$, $\mu_C = 0$, and $\mu_E = \delta$, we transformed each simulated experimental observation by adding the desired population mean. For each of the nine conditions, 10,000 replications were generated, and the bias of the eight indices was computed in the same way as for the simulations under the normality assumption.

Results

First, we show the results of bias and sampling variances assuming normal distributions, and next we focus on the results for nonnormal distributions.

Bias of the Estimators

Tables 6 and 7 show the bias of the different estimators calculated as the difference between the mean of each estimator over the 10,000 replicates and the population standardized mean difference δ, assuming normal distributions. Therefore, positive values reflect an overestimation of δ, whereas negative values imply the opposite.

The standardized mean difference index d is the only effect-size estimator in our simulation study that was calculated on the continuous values before dichotomizing the dependent variable. So, it was included only for comparison purposes. As expected, the d index exhibited the best performance in all of the conditions (for $\delta = 0.2$, bias = 0.0005; for $\delta = 0.5$, bias = 0.0002; and for $\delta = 0.8$, bias = –0.0003).

The seven other effect-size indices represent different alternatives for estimating the population standardized mean difference when the dependent variable has been dichotomized. From these, the d_{Cox} index showed the best performance in most of the conditions (for $\delta = 0.2$, bias = 0.0025; for $\delta = 0.5$, bias = 0.0076; and for $\delta = 0.8$, bias = 0.0238), although with a very slight overestimation of δ. The d_{Probit} index performed second best (for $\delta = 0.2$, bias = 0.0084; for $\delta = 0.5$, bias = 0.0189; and for $\delta = 0.8$, bias = 0.0293), but it slightly overestimated δ, its overestimation rising as δ increased. In fact, as shown in Tables 6 and 7, all values for the d_{Probit} index were positive. The d_{bis} index also overestimated δ (for $\delta = 0.2$, bias = 0.0103; for $\delta = 0.5$, bias = 0.0339; and for $\delta = 0.8$, bias = 0.0885). Again the overestimation increased with δ, this trend being more pronounced than in the case of the d_{Probit} index. Fourth, the d_{HH} index showed a slight underestimation that also increased with δ (for $\delta = 0.2$, bias = –0.0149; for $\delta = 0.5$, bias = –0.0367; and for $\delta = 0.8$, bias = –0.0506).

Finally, as expected, the d_p, d_ϕ, and d_{asin} indices showed a clear systematic underestimation (d_p: for $\delta = 0.2$, bias = –0.0334; for $\delta = 0.5$, bias = –0.0837; and for $\delta = 0.8$, bias = –0.1313. d_ϕ: for $\delta = 0.2$, bias = –0.0359; for $\delta = 0.5$, bias = –0.0854; and for $\delta = 0.8$, bias = –0.1415. d_{asin}: for $\delta = 0.2$, bias = –0.0357; for $\delta = 0.5$, bias = –0.0938; and for $\delta = 0.8$, bias = –0.1602) that increased with δ. In fact, for all of the manipulated conditions, these indices showed negative biases, their magnitude being clearly larger than those of the other indices. An example illustrates the pattern of the results: With $\delta = 0.5$, cutpoint $Y_c = 0.25$, and $n_E = n_C = 30$, the bias of the different effect-size estimators was $\text{Bias}(d) = 0.0029$, $\text{Bias}(d_{Cox}) = 0.0071$, $\text{Bias}(d_{Probit}) = 0.0184$, $\text{Bias}(d_{bis}) = 0.0386$, $\text{Bias}(d_{HH}) = -0.0387$, $\text{Bias}(d_p) = -0.0791$, $\text{Bias}(d_\phi) = -0.0862$, and $\text{Bias}(d_{asin}) = -0.0922$.

Of the different manipulated factors, the magnitude of the parametric effect size δ was the only one with a clear influence on the bias of the estimators, with the exception of the unbiased d index. As δ increased, the overestimation of d_{Cox}, d_{Probit}, and d_{bis} increased, and the underestimation of d_{HH}, d_p, d_ϕ, and d_{asin} also increased; the magnitude of this trend being different for each of the indices. As a consequence, the discrepancies among the indices are more pronounced

Table 6
Bias of the Effect-Size Indices for δ = *0.2 and* δ = *0.8*

Y_c	N	n_E	n_C	d	d_p	d_ϕ	d_{asin}	d_{HH}	d_{Cox}	d_{Probit}	d_{bis}
						δ = 0.2					
.1	48	24	24	−.0007	−.0319	−.0355	−.0358	−.0144	−.0040	.0087	.0126
.1	48	32	16	−.0004	−.0297	−.0333	−.0331	−.0106	.0082	.0126	.0155
.1	60	30	30	.0006	−.0329	−.0358	−.0362	−.0159	.0024	.0075	.0109
.1	60	40	20	.0008	−.0341	−.0368	−.0369	−.0162	.0021	.0070	.0095
.1	60	48	12	−.0019	−.0367	−.0393	−.0371	−.0149	.0035	.0073	.0061
.1	100	50	50	.0003	−.0356	−.0372	−.0375	−.0187	−.0007	.0049	.0070
.1	100	66	34	.0045	−.0325	−.0342	−.0343	−.0149	.0035	.0092	.0109
.1	100	80	20	.0008	−.0336	−.0352	−.0344	−.0139	.0046	.0098	.0095
						δ = 0.8					
.1	48	24	24	.0040	−.1257	−.1399	−.1591	−.0359	.0400	.0395	.1076
.1	48	32	16	.0010	−.1097	−.1252	−.1671	−.0511	.0232	.0258	.1564
.1	60	30	30	−.0010	−.1335	−.1447	−.1641	−.0475	.0272	.0294	.0906
.1	60	40	20	−.0026	−.1144	−.1266	−.1672	−.0549	.0191	.0233	.1386
.1	60	48	12	−.0019	−.0986	−.1115	−.1688	−.0574	.0163	.0208	.1796
.1	100	50	50	−.0011	−.1419	−.1486	−.1680	−.0601	.0133	.0196	.0720
.1	100	66	34	−.0003	−.1267	−.1339	−.1721	−.0667	.0061	.0130	.1109
.1	100	80	20	.0048	−.1032	−.1109	−.1663	−.0606	.0128	.0202	.1584
.4	48	24	24	−.0014	−.1150	−.1294	−.1537	−.0564	.0174	.0298	.1064
.4	48	32	16	.0021	−.1155	−.1297	−.1513	−.0499	.0246	.0351	.1054
.4	60	30	30	−.0026	−.1236	−.1349	−.1588	−.0661	.0068	.0206	.0893
.4	60	40	20	.0044	−.1152	−.1266	−.1493	−.0521	.0222	.0350	.1014
.4	60	48	12	−.0025	−.1238	−.1347	−.1486	−.0451	.0299	.0383	.0870
.4	100	50	50	.0010	−.1315	−.1382	−.1612	−.0750	−.0031	.0138	.0719
.4	100	66	34	.0031	−.1280	−.1348	−.1574	−.0695	.0030	.0194	.0766
.4	100	80	20	−.0004	−.1286	−.1352	−.1541	−.0614	.0120	.0264	.0746
.7	48	24	24	−.0067	−.1289	−.1430	−.1623	−.0401	.0353	.0352	.1041
.7	48	32	16	−.0015	−.1465	−.1593	−.1501	−.0170	.0607	.0545	.0549
.7	60	30	30	.0000	−.1314	−.1427	−.1624	−.0450	.0299	.0319	.0954
.7	60	40	20	−.0021	−.1590	−.1690	−.1639	−.0389	.0367	.0346	.0365
.7	60	48	12	−.0045	−.1722	−.1813	−.1438	−.0145	.0635	.0535	.0024
.7	100	50	50	.0020	−.1403	−.1469	−.1666	−.0588	.0148	.0213	.0746
.7	100	66	34	.0009	−.1618	−.1678	−.1679	−.0558	.0181	.0223	.0296
.7	100	80	20	−.0014	−.1759	−.1814	−.1597	−.0338	.0422	.0400	−.0008

for δ = 0.8 than for δ = 0.2. This trend was especially pronounced for d_{bis}, its overestimation of δ being remarkably large, in particular with cutpoint Y_c = 0.1 and imbalanced sample sizes for which the larger sample was assigned to the larger mean (that in our simulation always belonged to the experimental population). The poorest performance of d_{bis} happened under the conditions with the largest distance between the cutpoint and the parameter δ, and unbalanced sample sizes. On the other hand, neither the total sample size, the n_E:n_C ratio, the cutpoint, nor the relationship between sample size and the experimental and control means seemed to clearly affect the bias of the other effect-size estimators.

Sampling Variance of the Estimators

To compare the variances of the different estimators, for every manipulated condition we calculated the mean squared difference between each estimator and the population effect size δ for the 10,000 replicates. Note that as the bias of the estimator increases, this mean squared difference differs from the variance obtained, taking into account the empirical mean of the 10,000 replicates. However, when the purpose of the different effect-size indices is to estimate the same parameter (the population standardized mean difference δ), the deviations of interest are those with respect to the true parameter, δ in our study.

Table 7
Bias of the Effect-Size Indices for δ = *0.5*

Y_c	N	n_E	n_C	d	d_p	d_ϕ	d_{asin}	d_{HH}	d_{Cox}	d_{Probit}	d_{bis}
.1	48	24	24	.0004	−.0799	−.0887	−.0942	−.0347	.0115	.0199	.0415
.1	48	32	16	.0021	−.0746	−.0837	−.0944	−.0362	.0099	.0190	.0526
.1	48	16	32	−.0033	−.0849	−.0933	−.0913	−.0260	.0210	.0269	.0317
.1	60	30	30	.0053	−.0809	−.0879	−.0935	−.0360	.0100	.0193	.0388
.1	60	40	20	.0027	−.0755	−.0827	−.0931	−.0368	.0091	.0193	.0493
.1	60	20	40	−.0017	−.0878	−.0945	−.0934	−.0314	.0151	.0225	.0267
.1	60	48	12	.0009	−.0758	−.0830	−.0951	−.0379	.0080	.0173	.0514
.1	60	12	48	−.0008	−.0903	−.0966	−.0840	−.0137	.0346	.0376	.0198
.1	100	50	50	−.0016	−.0901	−.0942	−.0995	−.0471	.0022	.0091	.0242
.1	100	66	34	.0028	−.0834	−.0876	−.0969	−.0448	.0004	.0119	.0357
.1	100	34	66	−.0034	−.0892	−.0933	−.0940	−.0386	.0072	.0176	.0229
.1	100	80	20	.0020	−.0808	−.0851	−.0969	−.0439	.0013	.0126	.0415
.1	100	20	80	.0034	−.0920	−.0958	−.0894	−.0279	.0190	.0271	.0170
.25	48	24	24	−.0083	−.0857	−.0944	−.1002	−.0455	.0004	.0098	.0316
.25	48	32	16	.0038	−.0735	−.0823	−.0872	−.0285	.0183	.0278	.0488
.25	48	16	32	−.0014	−.0752	−.0841	−.0889	−.0305	.0161	.0256	.0456
.25	60	30	30	.0029	−.0791	−.0862	−.0922	−.0387	.0071	.0184	.0386
.25	60	40	20	−.0031	−.0814	−.0884	−.0933	−.0385	.0073	.0181	.0355
.25	60	20	40	−.0016	−.0845	−.0914	−.0965	−.0421	.0034	.0142	.0319
.25	60	48	12	.0033	−.0180	−.0878	−.0887	−.0288	.0180	.0265	.0358
.25	60	12	48	−.0029	−.0842	−.0910	−.0913	−.0315	.0150	.0232	.0317
.25	100	50	50	−.0011	−.0898	−.0939	−.0995	−.0506	−.0060	.0069	.0224
.25	100	66	34	−.0021	−.0903	−.0944	−.0995	−.0500	−.0054	.0073	.0216
.25	100	34	66	.0035	−.0851	−.0893	−.0945	−.0440	.0012	.0139	.0286
.25	100	80	20	−.0014	−.0871	−.0912	−.0942	−.0412	.0044	.0159	.0257
.25	100	20	80	.0028	−.0852	−.0892	−.0921	−.0387	.0071	.0188	.0282
.4	48	24	24	.0019	−.0729	−.0819	−.0880	−.0270	.0199	.0283	.0521
.4	48	32	16	−.0004	−.0838	−.0923	−.0899	−.0246	.0225	.0285	.0328
.4	48	16	32	−.0029	−.0734	−.0825	−.0932	−.0346	.0116	.0208	.0552
.4	60	30	30	.0010	−.0811	−.0881	−.0936	−.0365	.0096	.0191	.0384
.4	60	40	20	−.0001	−.0873	−.0937	−.0928	−.0314	.0152	.0230	.0275
.4	60	20	40	.0055	−.0765	−.0837	−.0938	−.0377	.0082	.0184	.0479
.4	60	48	12	.0009	−.0896	−.0959	−.0829	−.0120	−.0364	.0391	.0208
.4	60	12	48	.0010	−.0778	−.0849	−.0969	−.0402	.0055	.0151	.0487
.4	100	50	50	.0007	−.0910	−.0951	−.1004	−.0479	−.0031	.0082	.0233
.4	100	66	34	−.0023	−.0963	−.1003	−.1009	−.0469	−.0019	.0086	.0135
.4	100	34	66	−.0030	−.0878	−.0920	−.1011	−.0494	−.0047	.0068	.0301
.4	100	80	20	.0009	−.0949	−.0987	−.0927	−.0318	.0147	.0228	.0132
.4	100	20	80	.0013	−.0829	−.0871	−.0989	−.0463	−.0013	.0101	.0390

Tables 8–13 show the sampling variances of the estimators with respect to δ. As expected, the *d* index was the most efficient of the indices because it was the only one using all of the quantitative information in the dependent variable (for δ = 0.2, variance = 0.0712; for δ = 0.5, variance = 0.0756; and for δ = 0.8, variance = 0.0770). In the remaining indices, the cost of dichotomization implies a loss of information and, as a consequence, a loss of accuracy in the estimators. In effect, all of the indices showed higher sampling variances than that of the *d* index. Among the seven effect-size indices for 2 × 2 tables, d_p, d_ϕ, and d_{asin} had the lowest variances, although they were the most biased [d_p: for δ = 0.2, Var(d_p) = 0.0770; for δ = 0.5, Var(d_p) = 0.0911; and for δ = 0.8, Var(d_p) = 0.1096. d_ϕ: for δ = 0.2, Var(d_ϕ) = 0.0749; for δ = 0.5, Var(d_ϕ) = 0.0895; and for δ = 0.8, Var(d_ϕ) = 0.1094. d_{asin}: for δ = 0.2, Var(d_{asin}) = 0.0750; for δ = 0.5, Var(d_{asin}) = 0.0864; and for δ = 0.8, Var(d_{asin}) = 0.1028]. Next, d_{HH} and d_{Cox} obtained variances larger than those of d_p, d_ϕ, and d_{asin} [d_{HH}: for δ = 0.2, Var(d_{HH}) = 0.0957; for δ =

Table 8
Empirical (and Theoretical) Sampling Variances of the Effect-Size Indices for $\delta = 0.2$

Y_c	N	n_E	n_C	d	d_p	d_ϕ	d_{asin}	d_{HH}	d_{Cox}	d_{Probit}	d_{bis}
.1	48	24	24	.0832	.0896	.0861	.0849	.1077	.1299	.1356	.1438
				(.0846)	(.0845)	(.0884)	(.0833)	(.1069)	(.1291)	(.1349)	(.1502)
.1	48	32	16	.0929	.1046	.1006	.1005	.1302	.1573	.1630	.1689
				(.0951)	(.0951)	(.0995)	(.0938)	(.1224)	(.1478)	(.1532)	(.1696)
.1	60	30	30	.0661	.0723	.0701	.0691	.0863	.1040	.1091	.1149
				(.0676)	(.0675)	(.0701)	(.0667)	(.0847)	(.1022)	(.1073)	(.1172)
.1	60	40	20	.0761	.0817	.0793	.0790	.1001	.1207	.1260	.1301
				(.0760)	(.0759)	(.0787)	(.0750)	(.0964)	(.1164)	(.1215)	(.1315)
.1	60	48	12	.1029	.1104	.1072	.1115	.1482	.1788	.1827	.1748
				(.1054)	(.1053)	(.1088)	(.1042)	(.1401)	(.1692)	(.1729)	(.1808)
.1	100	50	50	.0405	.0435	.0428	.0423	.0515	.0618	.0652	.0676
				(.0404)	(.0403)	(.0414)	(.0400)	(.0500)	(.0604)	(.0638)	(.0667)
.1	100	66	34	.0454	.0481	.0473	.0470	.0580	.0698	.0735	.0755
				(.0450)	(.0449)	(.0461)	(.0446)	(.0560)	(.0677)	(.0714)	(.0753)
.1	100	80	20	.0628	.0657	.0660	.0657	.0834	.1005	.1049	.1031
				(.0630)	(.0630)	(.0644)	(.0625)	(.0804)	(.0971)	(.1014)	(.1048)

Table 9
Empirical (and Theoretical) Sampling Variances of the Effect-Size Indices for $\delta = 0.5$ *and* $Y_c = .1$

Y_c	N	n_E	n_C	d	d_p	d_ϕ	d_{asin}	d_{HH}	d_{Cox}	d_{Probit}	d_{bis}
.1	48	24	24	.0845	.1043	.1017	.0949	.1191	.1426	.1450	.1788
				(.0868)	(.0862)	(.0957)	(.0833)	(.1137)	(.1373)	(.1396)	(.1837)
.1	48	32	16	.0969	.1200	.1166	.1074	.1354	.1621	.1653	.2143
				(.0974)	(.0968)	(.1070)	(.0938)	(.1270)	(.1533)	(.1564)	(.2099)
.1	48	16	32	.0979	.1142	.1115	.1094	.1458	.1758	.1743	.1896
				(.0973)	(.0967)	(.1063)	(.0938)	(.1348)	(.1627)	(.1612)	(.1983)
.1	60	30	30	.0692	.0843	.0829	.0778	.0950	.1133	.1156	.1402
				(.0694)	(.0688)	(.0757)	(.0667)	(.0900)	(.1086)	(.1110)	(.1409)
.1	60	40	20	.0796	.0967	.0948	.0878	.1074	.1282	.1317	.1679
				(.0778)	(0.772)	(.0846)	(.0750)	(.1001)	(.1209)	(.1241)	(.1609)
.1	60	20	40	.0790	.0941	.0926	.0907	.1174	.1409	.1409	.1502
				(.0777)	(.0771)	(.0841)	(.0750)	(.1057)	(.1276)	(.1277)	(.1521)
.1	60	48	12	.1057	.1302	.1273	.1211	.1552	.1858	.1881	.2271
				(.1071)	(.1067)	(.1149)	(.1042)	(.1434)	(.1732)	(.1752)	(.2152)
.1	60	12	48	.1079	.1187	.1167	.1257	.1733	.2104	.2039	.1840
				(.1071)	(.1065)	(.1139)	(.1042)	(.1602)	(.1935)	(.1855)	(.1985)
.1	100	50	50	.0408	.0529	.0528	.0503	.0556	.0645	.0666	.0785
				(.0414)	(.0411)	(.0445)	(.0400)	(.0528)	(.0637)	(.0658)	(.0791)
.1	100	66	34	.0466	.0584	.0581	.0548	.0612	.0715	.0743	.0918
				(.0461)	(.0457)	(.0493)	(.0446)	(.0581)	(.0701)	(.0728)	(.0885)
.1	100	34	66	.0468	.0580	.0578	.0562	.0660	.0780	.0796	.0861
				(.0461)	(.0557)	(.0491)	(0.446)	(.0605)	(.0731)	(0.744)	(.0859)
.1	100	80	20	.0636	.0798	.0790	.0749	.0883	.1044	.1080	.1299
				(.0641)	(.0637)	(.0677)	(.0625)	(.0822)	(.0992)	(.1026)	(.1209)
.1	100	20	80	.0620	.0715	.0711	.0729	.0945	.1136	.1130	.1051
				(.0641)	(.0636)	(.0673)	(.0625)	(.0896)	(.1082)	(.1075)	(.1143)

Table 10
Empirical (and Theoretical) Sampling Variances of the Effect-Size Indices for $\delta = 0.5$ *and* $Y_c = .25$

Y_c	N	n_E	n_C	d	d_p	d_ϕ	d_{asin}	d_{HH}	d_{Cox}	d_{Probit}	d_{bis}
.25	48	24	24	.0856	.1066	.1040	.0971	.1184	.1406	.1444	.1785
				(.0867)	(.0862)	(.0955)	(.0833)	(.1113)	(.1344)	(.1378)	(.1814)
.25	48	32	16	.0988	.1164	.1133	.1070	.1368	.1647	.1677	.2196
				(.0974)	(.0968)	(.1069)	(.0938)	(.1282)	(.1548)	(.1572)	(1.010)
.25	48	16	32	.0983	.1166	.1135	.1075	.1374	.1651	.1677	.1997
				(.0974)	(.0968)	(.1069)	(.0938)	(.1282)	(.1548)	(.1571)	(.2015)
.25	60	30	30	.0695	.0843	.0828	.0773	.0926	.1102	.1137	.1394
				(.0693)	(.0688)	(.0758)	(.0667)	(.0881)	(.1064)	(.1098)	(.1396)
.25	60	40	20	.0773	.0928	.0912	.0865	.1059	.1263	.1296	.1524
				(.0777)	(.0772)	(.0843)	(.0750)	(.1004)	(.1213)	(.1244)	(.1536)
.25	60	20	40	.0770	.0966	.0949	.0899	.1100	.1308	.1341	.1583
				(.0777)	(.0772)	(.0843)	(.0750)	(.1005)	(.1213)	(.1244)	(.1541)
.25	60	48	12	.1073	.1249	.1224	.1223	.1607	.1935	.1940	.2043
				(.1072)	(.1066)	(.1145)	(.1042)	(.1475)	(.1781)	(.1778)	(.2043)
.25	60	12	48	.1075	.1254	.1230	.1240	.1627	.1956	.1952	.2042
				(.1071)	(.1066)	(.1144)	(.1042)	(.1481)	(.1788)	(.1780)	(.2043)
.25	100	50	50	.0409	.0536	.0534	.0509	.0558	.0644	.0669	.0789
				(.0414)	(.0411)	(.0445)	(.0400)	(.0519)	(.0626)	(.0652)	(.0784)
.25	100	66	34	.0463	.0581	.0579	.0556	.0620	.0720	.0747	.0860
				(.0460)	(.0457)	(.0491)	(.0446)	(.0581)	(.0702)	(.0728)	(.0860)
.25	100	34	66	.0476	.0583	.0581	.0556	.0631	.0739	.0766	.0886
				(.0461)	(.0457)	(.0492)	(.0865)	(.0583)	(.0704)	(.0729)	(.0865)
.25	100	80	20	.0643	.0759	.0753	.0742	.0896	.1063	.1089	.1157
				(.0641)	(.0637)	(.0675)	(.0625)	(.0839)	(.1014)	(.1038)	(.1162)
.25	100	20	80	.0622	.0745	.0740	.0728	.0884	.1050	.1075	.1145
				(.0641)	(.0637)	(.0675)	(.0625)	(.0840)	(.1015)	(.1038)	(.1163)

0.5, $\text{Var}(d_{HH}) = 0.1080$; and for $\delta = 0.8$, $\text{Var}(d_{HH}) = 0.1204$. d_{Cox}: for $\delta = 0.2$, $\text{Var}(d_{Cox}) = 0.1154$; for $\delta = 0.5$, $\text{Var}(d_{Cox}) = 0.1290$; and for $\delta = 0.8$, $\text{Var}(d_{Cox}) = 0.1430$]. The fact that the d_{HH} index had a slightly lower variance than d_{Cox} was due to the different multiplier constants used in both formulas. Very close to the variance of d_{Cox} index was that of d_{Probit}, the latter being even lower than the former for $\delta = 0.8$ [for $\delta = 0.2$, $\text{Var}(d_{Probit}) = 0.1200$; for $\delta = 0.5$, $\text{Var}(d_{Probit}) = 0.1306$; and for $\delta = 0.8$, $\text{Var}(d_{Probit}) = 0.1369$]. The least efficient estimator was d_{bis}, especially for $\delta = 0.8$ [for $\delta = 0.2$, $\text{Var}(d_{bis}) = 0.1223$; for $\delta = 0.5$, $\text{Var}(d_{bis}) = 0.1496$; and for $\delta = 0.8$, $\text{Var}(d_{bis}) = 0.2096$]. The larger bias of d_{bis} with $\delta = 0.8$ together with the great heterogeneity of this index explain this result.

Several factors affected the variances of the estimators. First, there was a direct relationship between the magnitude of the parameter δ and the values of the empirical variances for all of the effect-size indices. This trend was due to the larger bias of the indices as δ increased, as shown in the previous section. As expected, the larger the sample size, the lower the empirical variances. The imbalance between the sample sizes of experimental and control groups also affected the variances. In particular, the variances increased as the imbalance increased. Neither the relationship between the sample sizes and the experimental and control means nor the cutpoint seemed to affect the variances in a clear way. An exception was the variance of d_{bis}, where for $\delta = 0.8$, $Y_c = 0.1$, and unbalanced sample sizes with the lowest sample size assigned to the control group, the largest variances occurred. For example, with $n_E = n_C = 24$, the empirical variance was 0.2690, whereas with $n_E = 32$ and $n_C = 16$, the variance was 0.3927 (Table 12). The higher variance is a consequence of the greater overestimation of d_{bis} under these conditions.

Another objective of our simulation study was to examine the sampling variance formulas derived from statistical theory for the effect-size indices. The performance of some of the formulas has never been tested yet is crucial for their role in weighting every single effect size by its inverse variance. Tables 8–13

Table 11
Empirical (and Theoretical) Sampling Variances of the Effect-Size Indices for $\delta = 0.5$ *and* $Y_c = .4$

Y_c	N	n_E	n_C	d	d_p	d_ϕ	d_{asin}	d_{HH}	d_{Cox}	d_{Probit}	d_{bis}
.4	48	24	24	.0878	.1077	.1048	.0972	.1238	.1492	.1514	.1921
				(.0869)	(.0863)	(.0962)	(.0833)	(.1142)	(.1379)	(.1399)	(.1913)
.4	48	32	16	.0981	.1122	.1095	.1078	.1436	.1733	.1718	.1854
				(.0974)	(.0966)	(.1062)	(.0938)	(.1350)	(.1628)	(.1614)	(.1970)
.4	48	16	32	.0976	.1211	.1177	.1081	.1366	.1638	.1670	.2221
				(.0973)	(.0969)	(.1072)	(.0938)	(.1273)	(.1537)	(.1566)	(.2188)
.4	60	30	30	.0696	.0832	.0818	.0767	.0931	.1110	.1135	.1379
				(.0693)	(.0688)	(.0756)	(.0667)	(.0898)	(.1085)	(.1109)	(.1407)
.4	60	40	20	.0769	.0910	.0896	.0874	.1124	.1349	.1351	.1457
				(.0778)	(.0771)	(.0840)	(.0750)	(.1053)	(.1271)	(.1275)	(.1519)
.4	60	20	40	.0774	.0949	.0931	.0861	.1047	.1249	.1284	.1635
				(.0777)	(.0772)	(.0846)	(.0750)	(.0999)	(.1207)	(.1241)	(.1593)
.4	60	48	12	.1072	.1185	.1165	.1264	.1751	.2127	.2053	.1843
				(.1071)	(.1065)	(.1140)	(.1042)	(.1610)	(.1944)	(.1861)	(.1988)
.4	60	12	48	.1061	.1308	.1280	.1217	.1557	.1863	.1887	.2262
				(.1071)	(.1067)	(.1149)	(.1042)	(.1432)	(.1730)	(.1751)	(.2140)
.4	100	50	50	.0412	.0537	.0535	.0511	.0567	.0657	.0677	.0796
				(.0415)	(.0411)	(.0445)	(.0400)	(.0528)	(.0638)	(.0658)	(.0792)
.4	100	66	34	.0466	.0579	.0578	.0563	.0647	.0755	.0771	.0830
				(.0460)	(.0456)	(.0490)	(.0446)	(.0603)	(.0728)	(.0743)	(.0853)
.4	100	34	66	.0461	.0590	.0588	.0556	.0616	.0715	.0741	.0910
				(.0460)	(.0457)	(.0492)	(.0446)	(.0580)	(.0701)	(.0728)	(.0882)
.4	100	80	20	.0659	.0747	.0742	.0758	.0975	.1168	.1164	.1094
				(.0641)	(.0636)	(.0673)	(.0625)	(.0894)	(.1079)	(.1074)	(.1143)
.4	100	20	80	.0646	.0808	.0800	.0761	.0894	.1054	.1090	.1306
				(.0641)	(.0637)	(.0677)	(.0625)	(.0822)	(.0992)	(.1026)	(.1208)

show (in parentheses) the results of applying the formulas in each of the manipulated conditions, the values in the table being the means across the 10,000 replicates. As expected, the d index showed optimal sampling variance estimation, being very close to both the empirical variance and the mean variance from the formula (Equation 6) in all of the conditions. The formulas for the variances for d_{HH}, d_{Cox}, and d_{Probit} indices (Equations 17, 19, and 21, respectively) slightly underestimated the empirical variances. For example, with $\delta = 0.5$, cutpoint $Y_c = 0.25$, $n_E = n_C = 30$, the empirical variances for d_{HH}, d_{Cox}, and d_{Probit} were 0.0926, 0.1102, and 0.1137, respectively, whereas the mean values obtained with the formulas were 0.0881, 0.1064, and 0.1098 (see Table 10). It can be assumed that the discrepancies are negligible and also due to the slight bias of the estimators that was noted in the previous section.

The formula for the variance of d_{bis} (Equation 24) estimated the empirical variance well for $\delta = 0.2$ and $\delta = 0.5$, in all conditions. However, with $\delta = 0.8$ it suffered a more irregular performance; in particular, with $Y_c = 0.1$ and unbalanced sample sizes, the variance obtained by the formula was clearly larger than the empirical value. For example, for $\delta = 0.8$, $Y_c = 0.1$, $n_E = 32$ and $n_C = 16$, the mean variance with the formula was 0.5179, whereas the empirical variance was 0.3927 (see Table 12).

Finally, the formulas of the variances for d_p, d_ϕ, and d_{asin} (Equations 9, 12, and 14, respectively) systematically underestimated the empirical variances, the magnitude of such a deviation being higher as δ increased. This result is explained by the bias of these indices as estimators of δ, which also increased with the magnitude of δ, as noted in the previous section.

Nonnormal Distributions

Although the main focus of this article was testing the performance of different effect-size indices under the normality assumption, we have included a few conditions in which this assumption is not met to tentatively explore the robustness of these indices to violations of the normal distribution. There are many different ways for data to be nonnormally distributed,

Table 12
Empirical (and Theoretical) Sampling Variances of the Effect-Size Indices for $\delta = 0.8$ *and* $Y_c = .1$ *and* $Y_c = .4$

Y_c	N	n_E	n_C	d	d_p	d_ϕ	d_{asin}	d_{HH}	d_{Cox}	d_{Probit}	d_{bis}
.1	48	24	24	.0907	.1275	.1266	.1142	.1448	.1751	.1645	.2690
				(.0910)	(.0892)	(.1100)	(.0833)	(.1337)	(.1615)	(.1521)	(.2917)
.1	48	32	16	.1011	.1493	.1467	.1258	.1488	.1772	.1730	.3927
				(.1015)	(.1001)	(.1226)	(.0938)	(.1396)	(.1685)	(.1646)	(.5179)
.1	60	30	30	.0725	.1044	.1047	.0960	.1110	.1321	.1261	.1989
				(.0726)	(.0711)	(.0863)	(.0667)	(.1040)	(.1255)	(.1200)	(.2056)
.1	60	40	20	.0786	.1185	.1176	.1041	.1146	.1353	.1334	.2700
				(.0810)	(.0798)	(.0963)	(.0750)	(.1096)	(.1323)	(.1304)	(.2510)
.1	60	48	12	.1108	.1674	.1644	.1416	.1661	.1970	.1955	.4222
				(.1104)	(.1096)	(.1280)	(.1042)	(.1510)	(.1823)	(.1804)	(.3546)
.1	100	50	50	.0436	.0703	.0712	.0689	.0653	.0747	.0727	.1094
				(.0434)	(.0424)	(.0505)	(.0400)	(.0602)	(.0727)	(.0707)	(.1087)
.1	100	66	34	.0492	.0795	.0799	.0767	.0716	.0812	.0811	.1526
				(.0480)	(.0471)	(.0558)	(.0446)	(.0635)	(.0766)	(.0764)	(.1279)
.1	100	80	20	.0670	.1036	.1032	.0946	.0962	.1120	.1135	.2342
				(.0661)	(.0654)	(.0749)	(.0625)	(.0866)	(.1045)	(.1056)	(.1737)
.4	48	24	24	.0901	.1256	.1244	.1093	.1270	.1500	.1493	.2659
				(.0909)	(.0894)	(.1108)	(.0833)	(.1204)	(.1453)	(.1441)	(.2901)
.4	48	32	16	.1044	.1390	.1375	.1239	.1523	.1816	.1779	.2872
				(.1015)	(.0999)	(.1216)	(.0938)	(.1398)	(.1688)	(.1648)	(.3073)
.4	60	30	30	.0725	.1059	.1058	.0956	.1044	.1210	.1212	.1996
				(.0726)	(.0712)	(.0870)	(.0667)	(.0948)	(.1144)	(.1143)	(.1998)
.4	60	40	20	.0818	.1139	.1134	.1030	.1209	.1433	.1419	.2240
				(.0811)	(.0797)	(.0961)	(.0750)	(.1093)	(.1320)	(.1303)	(.2235)
.4	60	48	12	.1105	.1464	.1454	.1424	.1804	.2164	.2085	.2688
				(.1104)	(.1091)	(.1256)	(.1042)	(.1632)	(.1971)	(.1878)	(.2689)
.4	100	50	50	.0430	.0685	.0693	.0664	.0619	.0680	.0688	.1082
				(.0434)	(.0425)	(.0508)	(.0400)	(.0555)	(.0671)	(.0678)	(.1052)
.4	100	66	34	.0491	.0744	.0750	.0719	.0710	.0799	.0806	.1208
				(.0480)	(.0471)	(.0556)	(.0446)	(.0625)	(.0754)	(.0759)	(.1135)
.4	100	80	20	.0649	.0889	.0894	.0874	.0965	.11224	.1113	.1436
				(.0660)	(.0651)	(.0736)	(.0625)	(.0908)	(.1096)	(.1084)	(.1432)

so, we have included only three possible nonnormal distributions resulting from combining different values of skewness and kurtosis. Table 14 shows the bias of the different effect-size indices in respect to δ for the nine simulated conditions of population distribution δ and cutpoint Y_c. As a point of reference, the table also includes the bias of the effect indices assuming normal distributions (skewness = 0 and kurtosis = 0).

The simulated shapes of the distributions made it possible to explore the effect on the bias of increasing the degree of skewness (values of 0, 0.5, and 0.75) while the kurtosis was kept constant with value 0 and to compare these results with the ones in a more realistic condition in which the values of skewness and kurtosis differ from 0 (skewness = 1.75 and kurtosis = 3.75). The most robust of the indices was *d,* which was scarcely affected by changes in the shape of the distributions, its estimated bias being very close to zero. On the other hand, the negative bias of d_p, d_ϕ, d_{asin}, and d_{HH} indices was greater as the skewness increased, and the slightly positive bias of d_{Cox}, d_{Probit}, and d_{bis} indices seemed to first decrease and then change to a greater and greater negative bias as the skewness increased. As in the case of normal distributions, the magnitude of the population effect size affected the bias of the effect indices, the latter increasing as the former also increased.

Comparing the performance of the indices in the adverse (slightly nonnormal) conditions, we found that the d_p, d_ϕ, and d_{asin} indices performed much as they did under normal distributions, showing the largest bias, with systematically negative values [average bias across the nine nonnormal conditions: Bias(d_p)

Table 13
Empirical (and Theoretical) Sampling Variances of the Effect-Size Indices for $\delta = 0.8$ *and* $Y_c = .7$

Y_c	N	n_E	n_C	d	d_p	d_ϕ	d_{asin}	d_{HH}	d_{Cox}	d_{Probit}	d_{bis}
.7	48	24	24	.0903	.1298	.1289	.1159	.1451	.1746	.1644	.2762
				(.0908)	(.0892)	(.1099)	(.0833)	(.1335)	(.1612)	(.1521)	(.3024)
.7	48	32	16	.1022	.1291	.1291	.1286	.1785	.2191	.1973	.3139
				(.1015)	(.0993)	(.1185)	(.0938)	(.1705)	(.2059)	(.1831)	(.2564)
.7	60	30	30	.0755	.1071	.1072	.0982	.1155	.1380	.1380	.2131
				(.0726)	(.0711)	(.0866)	(.0667)	(.1045)	(.1262)	(.1204)	(.2182)
.7	60	40	20	.0811	.1106	.1114	.1090	.1420	.1712	.1556	.1678
				(.0810)	(.0791)	(.0931)	(.0750)	(.1305)	(.1576)	(.1432)	(.1938)
.7	60	48	12	.1128	.1340	.1347	.1544	.2059	.2526	.2234	.1802
				(.1104)	(.1083)	(.1221)	(.1042)	(.2074)	(.2504)	(.2137)	(.2282)
.7	100	50	50	.0432	.0703	.0712	.0688	.0653	.0750	.0732	.1109
				(.0434)	(.0424)	(.0506)	(.0400)	(.0601)	(.0726)	(.0708)	(.1092)
.7	100	66	34	.0474	.0751	.0762)	.0741	.0778	.0906	.0854	.0942
				(.0480)	(.0468)	(.0544)	(.0446)	(.0721)	(.0870)	(.0821)	(.1075)
.7	100	80	20	.0668	.0916	.0927	.0955	.1265	.1532	.1361	.1072
				(.0660)	(.0648)	(.0720)	(.0625)	(.1155)	(.1394)	(.1235)	(.1314)

$= -0.1251$, Bias(d_ϕ) $= -0.1314$, and Bias(d_{asin}) $= -0.1382$]. Next, the d_{HH} index also showed a negative bias, although systematically lower than that of the previous indices, Bias(d_{HH}) $= -0.0832$. And finally, although with a more irregular pattern, d_{Cox}, d_{Probit}, and d_{bis} indices achieved the best performance, with the smallest values of bias [average bias: Bias(d_{Cox}) $= -0.0418$, Bias(d_{Probit}) $= -0.0352$, and Bias(d_{bis}) $= -0.0127$].

Discussion

The purpose of this article was to examine the performance of different effect-size indices that quantify the results of a 2 × 2 table in the d metric. These indices are very useful in meta-analysis, in particular when the meta-analyst finds that studies contain a mixture of continuous and dichotomized dependent variables. In this article we have carried out a Monte

Table 14
Bias of the Effect-Size Indices for $\delta = 0.2$ *and* $\delta = 0.8$ *Under Nonnormal Distributions*

Skew	Kurtosis	d	d_p	d_ϕ	d_{asin}	d_{HH}	d_{Cos}	d_{Probit}	d_{bis}
				$\delta = 0.2$; $Y_c = .1$					
0	0	.0006	−.0329	−.0358	−.0362	−.0159	.0024	.0075	.0109
0.5	0	−.0031	−.0431	−.0458	−.0460	−.0266	−.0094	−.0048	−.0019
0.75	0	−.0053	−.0612	−.0635	−.0637	−.0454	−.0301	−.0265	−.0239
1.75	3.75	.0025	−.0610	−.0633	−.0532	−.0399	−.0240	−.0226	−.0206
				$\delta = 0.5$; $Y_c = .25$					
0	0	.0029	−.0791	−.0862	−.0922	−.0387	.0071	.0184	.0386
0.5	0	−.0016	−.1052	−.1119	−.1167	−.0655	−.0224	−.0121	.0049
0.75	0	.0029	−.1388	−.1448	−.1487	−.0999	−.0602	−.0516	−.0372
1.75	3.75	.0107	−.1434	−.1494	−.1519	−.0913	−.0507	−.0479	−.0356
				$\delta = 0.8$; $Y_c = .4$					
0	0	−.0026	−.1236	−.1349	−.1588	−.0661	.0068	.0206	.0893
0.5	0	−.0003	−.1570	−.1679	−.1885	−.0998	−.0303	−.0172	.0427
0.75	0	−.0007	−.2121	−.2220	−.2374	−.1547	−.0906	−.0791	−.0329
1.75	3.75	.0207	−.2039	−.2139	−.2278	−.1239	−.0584	−.0551	−.0098

Carlo simulation to examine the bias, the empirical variance, and the adjustment of the formulas for estimating the sampling variance of seven different effect-size indices that can be obtained from a 2 × 2 table.

Two of these indices assume normal distributions (d_{Probit} and d_{bis}), two others are based on logistic distributions (d_{Cox} and d_{HH}), one applies the arcsine transformation (d_{asin}), and the remaining two (d_p and d_ϕ) apply the standardized-mean-difference formula without taking into account the dichotomization. We assumed normal and nonnormal distributions in the simulation. Under normal distributions, d_{Probit} and d_{bis} indices have an advantage. Including a few conditions with nonnormal distributions helped us to tentatively explore the robustness of these effect indices.

The best effect-size index should be the least biased and have a formula derived from statistical theory to properly estimate the sampling variance of the index. Under the normal-distribution assumption, our results show that of the seven indices compared, d_{Cox} is practically unbiased as an estimator of the population standardized mean difference δ, showing estimates very close to those of d index. Further, its good performance was not altered by the manipulated factors in the simulation. As expected, the sampling variances of the indices that take into account the effect of dichotomizing the outcome (d_{Cox}, d_{Probit}, d_{HH}, and d_{bis}) were larger than that of d, because dichotomizing variables has a cost in terms of accuracy. Although very close to d_{Cox}, the d_{Probit} index slightly overestimated the parameter δ, its variance also being very similar to that of d_{Cox}. In any case, the performance of both indices, d_{Cox} and d_{Probit}, is very similar for practical purposes.

The d_{bis} index presents a good performance with a low or medium magnitude in the parametric effect size but remarkably overestimates population standardized mean differences of a large magnitude. Further, under these conditions the formula for estimating the sample variance of d_{bis} (Equation 24) also overestimates the empirical variance. Therefore, d_{bis} should be used only when the meta-analyst can safely assume that the population effect sizes are of a small or medium magnitude.

The d_{HH} index slightly underestimated the population effect size as expected, although its sampling variance (Equation 17) performed well. The underestimation occurred because the d_{HH} index rescales the log odds ratio by dividing it by 1.81, which is the standard deviation of the logistic distribution. On the other hand, d_{Cox} rescales the log odds ratio by dividing it by 1.65, which produces estimates closer to the normal distribution. As a consequence, if the meta-analyst can assume normality, d_{Cox} and d_{Probit} are preferable to the other indices examined here.

The d_p, d_ϕ, and d_{asin} indices had the poorest performance, systematically underestimating the population effect size, and their formulas for computing the sampling variances (Equations 9, 12, and 14, respectively) also slightly underestimated the empirical variances. However, the sampling variances of these three indices were clearly lower than those of the d_{Cox}, d_{Probit}, d_{HH}, and d_{bis} indices. This is because the latter indices correct the artifact of dichotomizing continuous outcomes, introducing a multiplier constant that increases the sampling variance of the estimator, whereas the d_p, d_ϕ, and d_{asin} indices do not correct the negative bias due to dichotomization. As a consequence, use of these indices is ill-advised in all conditions because of their systematic negative bias. In summary, under the normal-distribution assumption, d_{Cox} and d_{Probit} offer the best results and would be the indices of choice when the meta-analyst can reasonably assume such a distribution.

Although the performance of the indices under nonnormal distributions was beyond the scope of this article, we added in our simulation a few conditions in which the degree of skewness was manipulated. Taking into account that the manipulated conditions were not representative of the universe of possible nonnormal distributions, we found that as the degree of skewness increased, the negative bias of the d_p, d_ϕ, d_{asin}, and d_{HH} indices also increased. On the contrary, the positive bias of the d_{Cox}, d_{Probit}, and d_{bis} indices decreased and then changed to a negative bias greater as skewness increased. Comparing the performance of the indices, we found that d_{bis}, d_{Probit}, and d_{Cox} indices showed the lowest bias. Next, the d_{HH} index showed a systematic negative bias of a larger magnitude, and finally, the d_p, d_ϕ, and d_{asin} indices continued showing the worst performance, with a systematic negative bias, the same as under normal distributions.

A limitation for generalizing the results of our simulation study was to assume normal distributions in most of conditions. However, primary studies in social and behavioral sciences routinely apply parametric statistical tests that imply normality. Therefore, the purpose of our article was to offer initial empirical evidence of the performance of several effect-size indices under the most usually assumed conditions. Future research is needed to examine the performance of

the effect-size indices under other distributions, where the index of choice might be different than d_{Cox} or d_{Probit}. For example, it is important to know the performance of these indices under very skewed and/or heteroskedastic distributions and extreme success (or failure) proportions that can appear in real data.[2]

With nonnormal distributions, the standardized mean difference probably will not be the best population effect-size index to reveal the effect magnitude between two populations and, as a consequence, the effect indices tested here would not work well either. In these cases it should be more advisable to apply other effect indices, such as the nonparametric effect sizes proposed by Kraemer and Andrews (1982) and Hedges and Olkin (1985). One problem is that these indices require the individual data for the two samples, and this information is rarely reported in the studies. Therefore, the meta-analyst will face serious limitations when applying these indices.

In any case, the meta-analyst should include a moderator variable for testing whether there are differences between the *d* indices obtained from the studies with continuous outcomes and the ones with dichotomized outcomes (Lipsey & Wilson, 2001). If there are no differences between the two *d* metrics, it should be advisable to maintain all of the studies in a same meta-analysis, but if there are differences between them, a better solution could be not mixing studies with continuous and dichotomized outcomes but doing two separate meta-analyses, one for the studies with outcomes measured continuously (e.g., using *d* indices) and another one for the studies with dichotomized outcomes using some of the effect-size indices presented here, or also odds ratios, risk ratios, or risk differences.

It is important to note that the focus of our article was how to obtain an effect-size index in the *d* metric from a single study when the outcome has been dichotomized, to integrate it in a meta-analysis in which some of the studies reported their results continuously and others reported them dichotomized. A different, but related, problem occurs when the meta-analyst finds studies with continuous outcomes and studies with true dichotomies. In this case there are two different sets of parameters: the population means (and standard deviations) for the continuous outcomes and the population proportions for the dichotomous outcomes. If the meta-analyst wants to mix all of the studies in the same meta-analysis, the effect-size indices treated here for dichotomized outcomes could be tentatively used on the true dichotomies, with the caution of including a moderator variable for testing possible differences between the two *d* metrics.

In summary, although more research is needed to know the performance of these effect indices under other conditions and distribution assumptions, our results can help, on the one hand, to select the effect-size index in future meta-analyses that include dichotomous and/or dichotomized variables and, on the other hand, to assess possible underestimations of the population effect size in past meta-analyses that have used such effect-size indices as d_p, d_{asin}, d_ϕ, or ϕ.

[2] Extreme proportions would happen when the cutpoint is far from both the experimental and control means, μ_E and μ_C. In our simulations we have found no effect of the cutpoint on the bias and sampling variance of the effect-size indices. However, our simulations only included cutpoints placed between the two means, giving rise to not very extreme proportions. So, to explore the possible influence of the cutpoint on the bias of the effect-size indices, we have carried out additional simulations moving the cutpoint to extreme positions. In particular, assuming normality, $n_E = n_C = 30$, and $\delta = 0.5$ (μ_E 0.5, $\mu_C = 0$, and $\sigma = 1$), we have examined the bias of the d_{Cox}, d_{Probit}, and d_{HH} indices, fixing the cutpoint in such extreme values as $Y_c = 1.3$, 1.4, 1.6, 1.8, 2.0, and 2.1. The results showed a negative relationship between the cutpoint and the bias of the estimators, changing from an overestimation to an underestimation of the parametric effect size, δ [Bias(d_{Cox}) = 0.116, 0.136, 0.125, 0.106, 0.039, and –0.003; Bias(d_{Probit}) = 0.036, 0.039, 0.005, –0.029, –0.094, and –0.0131; Bias(d_{HH}) = 0.061, 0.078, 0.069, 0.052, –0.010, and –0.048, respectively]. The highest bias happened for $Y_c = 1.4$, and it was of 27.2%, 7.8%, and 15.6% for the d_{Cox}, d_{Probit}, and d_{HH} indices, respectively. For practical purposes, we have found a good performance for the three effect indices even when the cutpoint is so extreme that it could give rise to very small cell frequencies (for example, with $O_{1E} = 1$ and $O_{1C} = 0$). So, although more research is needed to examine the robustness of these effect-size indices, our results show that, in general, they are a good solution for translating into the *d* metric the results of studies with dichotomized outcomes.

References

Aptech Systems. (1992). *The Gauss system* (*Version 3.0*) [Computer software]. Kent, WA: Author.

Becker, G., & Thorndike, R. L. (1988). The biserial-phi correlation coefficient. *Journal of Psychology, 122,* 523–526.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine, 19,* 3127–3131.

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

Cooper, H. (1998). *Integrating research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Cooper, H., & Hedges, L. V. (Eds.) (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Cox, D. R. (1970). *Analysis of binary data.* New York: Chapman & Hall/CRC.

Deeks, J. J., & Altman, D. G. (2001). Effect measures for meta-analysis of trials with binary outcomes. In M. Egger, G. Davey Smith, & D. G. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (pp. 313–335). London: BMJ Books.

Dominici, F., & Parmigiani, G. (2000). Combining studies with continuous and dichotomous responses: A latent-variables approach. In D. K. Stangl & D. A. Berry (Eds.), *Meta-analysis in medicine and health policy* (pp. 105–125). New York: Marcel Dekker, Inc.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43,* 521–531.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

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

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

Haddock, C. K., Rindskopf, D., & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods, 3,* 339–353.

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin, 117,* 167–178.

Hasselblad, V., Mosteller, F., Littenberg, B., Chalmers, T. C., Hunink, M. G. M., Turner, J. A. et al. (1995). A survey of current problems in meta-analysis: Discussion from the Agency for Health Care Policy and Research Inter-PORT Work Group on Literature Review/Meta-analysis. *Medical Care, 33,* 202–220.

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

Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75,* 334–349.

Johnson, B. T. (1989). *DSTAT: Software for the meta-analysis review of research* [Computer software and manual]. Hillsdale, NJ: Erlbaum.

Johnson, B. T., & Eagly, A. H. (2000). Quantitative synthesis of social psychological research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 496–528). Cambridge, MA: Cambridge University Press.

Kraemer, H. C., & Andrews, G. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin, 91,* 404–412.

Laird, N., & Mosteller, F. (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care, 6,* 5–30.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

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. (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). Thousand Oaks, CA: Sage.

Sánchez-Meca, J., & Marín-Martínez, F. (2000). Testing the significance of a common risk difference in meta-analysis. *Computational Statistics & Data Analysis, 33,* 299–313.

Sánchez-Meca, J., & Marín-Martínez, F. (2001). Meta-analysis of 2 × 2 tables: Estimating a common risk difference. *Educational & Psychological Measurement, 61,* 249–276.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.

Stuart, A., & Ord, K. (1994). *Kendall's advanced theory of statistics. Vol. I: Distribution theory* (6th ed.). New York: Edward Arnold.

Thorndike, R. L. (1949). *Personnel selection.* New York: Wiley.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin.

Whitehead, A., Bailey, A. J., & Elbourne, D. (1999). Combining summaries of binary outcomes with those of continuous outcomes in a meta-analysis. *Journal of Biopharmaceutical Statistics, 9,* 1–16.

Received August 30, 2002
Revision received March 25, 2003
Accepted July 18, 2003 ■