Meta-Analysis of 2 x 2 Tables: Estimating a Common Risk Difference

Julio Sánchez-Meca and Fulgencio Marín-Martínez

Educational and Psychological Measurement 2001; 61; 249
DOI: 10.1177/00131640121971220

META-ANALYSIS OF 2 × 2 TABLES: ESTIMATING A COMMON RISK DIFFERENCE

JULIO SÁNCHEZ-MECA AND FULGENCIO MARÍN-MARTÍNEZ
Universidad de Murcia

Five procedures for estimating a common risk difference in a set of independent 2 × 2 tables were assessed via Monte Carlo simulation in terms of their bias, efficiency, confidence level adjustment, and statistical power. The maximum likelihood estimator showed the best performance, very closely followed by Cochran's and Mantel-Haenszel's procedures. The conditional weighted estimator, d_{CW}, showed an irregular performance. The unweighted estimator, d_U, showed less efficiency and statistical power than that of the other procedures. As a consequence, the use of the d_{CW} and d_U estimators is not recommended. The implications of the results in the practice of meta-analysis are discussed.

Meta-analysis has become a very common research methodology in behavioral sciences. It can be defined as the quantitative analysis of the results of a set of studies about a given research topic (Cooper & Hedges, 1994; Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985). Meta-analytic techniques have also been applied in measurement contexts; for example, in *Educational and Psychological Measurement*, Vacha-Haase (1998) proposed "reliability generalization" meta-analytic methods, and Schmidt and Hunter (1977) proposed "validity generalization" meta-analytic methods. To carry out a meta-analysis, an effect size index has to be selected that represents the outcome of each study, and, based on this, analysis techniques are applied to achieve three main objectives: (a) to estimate the average effect of the studies, (b) to test whether the set of studies is homogeneous regarding the estimated common effect size, and (c) if the homogeneity hypothesis is not met, to test the influence of potential moderator variables that explain such heterogene-

This research was supported by Grant PB98-0401 from Ministerio de Educación y Cultura, Spain.

Educational and Psychological Measurement, Vol. 61 No. 2, April 2001 249-276

ity (Cooper, 1989; Rosenthal, 1991; Sánchez-Meca & Marín-Martínez, 1998; Shadish, 1996).

When an empirical study compares two groups (usually treatment vs. control) using a continuous outcome measure, the standardized mean difference is the effect size most often applied in meta-analyses of the behavioral sciences (Rosenthal, 1994). In the same situation with a dichotomous outcome (e.g., success vs. failure), the usual effect size indexes are the risk difference, the risk ratio, and the odds ratio (Fleiss, 1994; Rothman, 1986).

Dichotomous results are not infrequent in behavioral sciences. For example, in psychological treatment evaluation, the results can be measured as cured versus ill or respond versus not respond to the treatment. In particular, in the treatment of tobacco addiction, it is very common to record the dichotomous results abstinence versus nonabstinence; also, in correctional treatment of delinquents, one of the most critical outcomes is the recidivism versus nonrecidivism into crime (e.g., Andrews et al., 1990; Whitehead & Lab, 1989).

Special procedures have to be applied when the purpose of a meta-analysis is to integrate a set of studies in which the results are summarized as 2×2 tables. Although psychological literature has scarcely paid attention to the integration of 2×2 tables (as an exception, see Haddock, Rindskopf, & Shadish, 1998), this topic has been widely developed in the health sciences (Greenland, 1987; Hasselblad & McCrory, 1995; Laird & Mosteller, 1990; Petitti, 2000).

Given a set of k independent studies with sample sizes N_i, $i = 1, 2, \ldots k$, each composed of two groups of subjects, a proportion q_i assigned to treatment, $q_i' = 1 - q_i$ assigned to control, and with observed success proportions p_{T_i} and p_{C_i}, the risk difference in study i is defined as $d_i = p_{T_i} - p_{C_i}$, the risk ratio as $rr_i = p_{T_i}/p_{C_i}$, and the odds ratio as $or_i = p_{T_i}(1 - p_{C_i})/p_{C_i}(1 - p_{T_i})$.

Each of these effect size measures has its advantages and disadvantages (Fleiss, 1994). The focus of the present study is the risk difference. In comparison to other, more complicated indexes, the risk difference is easy to interpret because it is the natural measure of the gain of one treatment with respect to another. For example, in life-threatening situations, risk difference can be translated into the number of people cured or the number of people who died and therefore has a direct implication for benefit (Hasselblad et al., 1995). Risk difference has an approximate normal distribution even with low sample sizes in the studies, and it is well defined even when some of the proportions are 0 or 1 (Laird & Mosteller, 1990).

Nevertheless, this index also has its problems. The possible values of risk difference when the success proportions are close to 0.5 are greater than when they are in the extremes (Fleiss, 1994). For example, a risk difference of

0.1 between two proportions is a large effect if $p_{C_i} = 0.02$, a modest-size effect if $p_{C_i} = 0.5$, and impossible if $p_{C_i} = 0.95$ (Laird & Mosteller, 1990). Moreover, if p_{T_i} and/or p_{C_i} values vary through the studies, d_i values will also vary, giving the appearance of heterogeneity owing to the mathematical constraints imposed on proportions rather than to substantive reasons (Fleiss, 1994). The same situation may occur when control group proportions, p_{C_i}, are very close to 0 or 1 (DerSimonian & Laird, 1986; Whitehead & Whitehead, 1991).

Here we assume complete homogeneity, that is, that the probability of success in all treatment groups is π_T and in all control groups is π_C; thus, the risk difference in each study is $\delta = \pi_T - \pi_C$. In other words, we assume that there are no moderator variables affecting the individual studies' variability. We also assume a fixed-effects model in which the variability among d_i values is only due to within-studies variance. Thus, the k independent studies come from the same population with a common population risk difference δ. Therefore, inferences can only be made to studies similar to those included in the meta-analysis, that is, to studies belonging to the same population (Petitti, 2000). Under these conditions, the risk difference is an advisable effect size index to summarize the results of a set of k 2×2 tables.

When the complete homogeneity of effect size assumption is met, the main purpose of a meta-analysis is to estimate δ, to estimate its standard error, and to compute a confidence interval estimate of δ. Several estimators of a common risk difference have been proposed in the literature. Assuming large sample sizes, n_{T_i} and n_{C_i}, in the individual studies and a high number of studies, k, these estimates follow known distributions. However, in multiple real situations, meta-analyses are applied to a few studies with low sample sizes. Moreover, such factors as the imbalance between sample sizes, n_{T_i} and n_{C_i}, or the relationship between sample sizes and treated and control groups can affect the performance of the different estimators. The behavior of the different estimators through these conditions can help meta-analysts to decide which estimator to use in a particular meta-analysis.

Although some interest has been shown in comparing the performance of different common risk differences under the random-effects model (Emerson, Hoaglin, & Mosteller, 1993, 1996), less interest has been shown in the fixed-effects model. In this study we compared the bias, the efficiency, the adjustment of the confidence level, and the statistical power of five estimators of a common risk difference: (a) the maximum likelihood estimator, d_{ML} (e.g., Laird & Mosteller, 1990); (b) the common risk difference estimator with conditional weights, d_{CW} (e.g., Shadish & Haddock, 1994); (c) the Cochran common risk difference estimator, d_C (Cochran, 1954); (d) the Mantel-Haenszel common risk difference estimator, d_{MH} (Mantel & Haenszel, 1959); and (e) the unweighted common risk difference estimator, d_U.

Procedures for Estimating a Common Risk Difference

Maximum Likelihood Estimator

Consider a set of k two-group designs, each of them yielding an estimate, $d_i = p_{T_i} - p_{C_i}$, of the population common risk difference, $\delta = \pi_T - \pi_C$. The maximum likelihood estimate, d_{ML}, of δ is (Goldman & Feinstein, 1979):

$$d_{ML} = p_T - p_C, \tag{1}$$

where $p_T = x_T/NQ$ and $p_C = x_C/NQ'$ are the maximum likelihood estimators of π_T and π_C, respectively; x_T and x_C are the total numbers of successes in the k treatment groups and in the k control groups, respectively; $NQ = \Sigma_i N_i q_i$ and $NQ' = \Sigma_i N_i q'_i$, N being the total number of subjects seen in all the k studies and N_i being the total sample size of each study. Under the homogeneity assumptions, d_{ML} is approximately normally distributed, an unbiased estimator of δ, with variance σ^2_{ML}:

$$\sigma^2_{ML} = \frac{\pi_T(1-\pi_T)}{NQ} + \frac{\pi_C(1-\pi_C)}{NQ'}. \tag{2}$$

One could estimate the variance by substituting the maximum likelihood estimators of π_T and π_C in the above equation:

$$S^2_{ML} = \frac{p_T(1-p_T)}{NQ} + \frac{p_C(1-p_C)}{NQ'}. \tag{3}$$

This estimator would be slightly biased because when np has a binomial distribution with parameters n and π, $E[p(1-p)] = \pi(1-\pi)(n-1)/n < \pi(1-\pi)$. However, because with maximum likelihood estimation the number of successes is cumulated over all the k studies, this bias is likely to be negligible in a meta-analysis. An approximate $100(1-\alpha)\%$ level confidence interval estimate of δ is given by:

$$d_{ML} \pm |z_{\alpha/2}|\, S_{ML}. \tag{4}$$

Unweighted Common Risk Difference Estimator

The maximum likelihood estimator of a common risk difference is not a weighted average of the effect size estimators from the individual studies. Another unweighted estimator of a common risk difference is obtained averaging the individual risk differences of the studies, $d_i = p_{T_i} - p_{C_i}$:

$$d_U = \frac{\sum_i d_i}{k}. \tag{5}$$

The d_U estimator is also unbiased and follows an approximate t distribution with $k - 1$ degrees of freedom, with estimated variance:

$$S_U^2 = \frac{S_d^2}{k}, \tag{6}$$

and S_d^2 being the unbiased variance of the k individual risk differences, d_i:

$$S_d^2 = \frac{\sum_i (d_i - d_U)^2}{k-1}. \tag{7}$$

Thus, a confidence interval estimate of δ can be obtained by:

$$d_U \pm |_{\alpha/2} t_{k-1}| \, S_U. \tag{8}$$

One would expect that this confidence interval estimator would not work as well as the maximum likelihood estimate because the variance estimator, S_U^2, is less efficient ignoring all the information of binomial distributions used in Equation 3.

Conditional Common Risk Difference Estimator

With large sample sizes, $n_{T\,i}$ and $n_{C\,i}$, d_i estimates follow an approximate normal distribution, $N(\delta, \sigma_{d_i}^2)$, the d_i variance being:

$$\sigma_{d_i}^2 = \frac{\pi_T(1-\pi_T)}{n_{T_i}} + \frac{\pi_C(1-\pi_C)}{n_{C_i}}. \tag{9}$$

A δ estimate can be obtained calculating a weighted mean, d_+, of the individual risk differences, d_i (Fleiss, 1981, p. 162):

$$d_+ = \frac{\sum_i W_i d_i}{\sum_i W_i}. \tag{10}$$

The optimal weight of each d_i, W_i, is obtained as the inverse of the variance, $\sigma_{d_i}^2$, that is, $W_i = 1/\sigma_{d_i}^2$. But d_i variance is unknown because it is a function of the sample sizes, $n_{T\,i}$ and $n_{C\,i}$, and the unknown population proportions, π_T and π_C. An estimate of $\sigma_{d_i}^2$ can be obtained substituting the sample proportions, $p_{T\,i}$ and $p_{C\,i}$, for π_T and π_C (Shadish & Haddock, 1994, p. 270):

$$S_{d_i}^2 = \frac{p_{T_i}(1-p_{T_i})}{n_{T_i}} + \frac{p_{C_i}(1-p_{C_i})}{n_{C_i}}, \tag{11}$$

and W_i is estimated via $w_i = 1/S_{d_i}^2$. Thus, an unbiased estimator of δ is given by:

$$d_{CW} = \frac{\sum_i w_i d_i}{\sum_i w_i}. \tag{12}$$

With large sample sizes, $n_{T\,i}$ and $n_{C\,i}$, the d_{CW} estimator follows an approximately normal distribution with estimated variance:

$$S^2_{CW} = \frac{1}{\sum_i w_i}. \tag{13}$$

Then, a confidence interval estimate of δ is given by:

$$d_{CW} \pm |z_{\alpha/2}|\, S_{CW}. \tag{14}$$

We have named this estimate "conditional weighted common risk difference" because d_{CW} and w_i are both a function of the sample proportions, $p_{T\,i}$ and $p_{C\,i}$, and this departure of independence can affect the estimation of δ. This estimator theoretically would not work as well as d_{ML} because it is less efficient.

A practical problem arises in calculating $S^2_{d_i}$ when $p_{T\,i} = p_{C\,i} = 0$, because in this situation $S^2_{d_i} = 0$. To produce a nonzero value of $S^2_{d_i}$, we adapted an adjustment proposed by Tukey (1977, chap. 15) that replaces a proportion $p = x/n$ by $p^* = (x + 1/6)/(n + 1/3)$, where x is the number of successes and is binomially distributed, $B(n, \pi)$. This adjustment must be only applied in calculating $S^2_{d_i}$.

Cochran Common Risk Difference Estimator

To avoid dependence between d_i and w_i, Cochran (1954, p. 443; 1983, pp. 105-106) proposed to estimate δ by:

$$d_C = \frac{\sum_i w_i^* d_i}{\sum_i w_i^*}, \tag{15}$$

where the individual weights, w_i^*, are simply a function of the sample sizes, $n_{T\,i}$ and $n_{C\,i}$:

$$w_i^* = \frac{n_{T_i} n_{C_i}}{n_{T_i} + n_{C_i}}. \tag{16}$$

Thus, dependence between d_i and w_i is removed. With large sample sizes, d_C is an unbiased estimator that follows an approximately normal distribution with estimated variance:

$$S^2_C = \frac{\sum_i w_i^* \bar{p}_i (1 - \bar{p}_i)}{(\sum_i w_i^*)^2}, \tag{17}$$

where

$$\bar{p}_i = \frac{x_{T_i} + x_{C_i}}{N_i}, \tag{18}$$

x_{T_i} and x_{C_i} being the success numbers in treated and control groups, respectively, in the study i, and $N_i = n_{T\,i} + n_{C\,i}$. A confidence interval estimate is obtained by:

$$d_C \pm |z_{\alpha/2}|\, S_C. \tag{19}$$

Mantel-Haenszel Common Risk Difference Estimate

Finally, Mantel and Haenszel (1959) proposed the same common risk estimator as Cochran but with a slight modification in the estimated variance to avoid the binomial bias. Thus, the Mantel-Haenszel estimator of δ is defined as in Equation 15: $d_{MH} = \Sigma_i w_i^* d_i / \Sigma_i w_i^*$, with w_i^* defined in Equation 16, and the estimated variance of d_{MH} being:

$$S_{MH}^2 = \frac{\sum_i w_i \bar{p}_i (1-\bar{p}_i)}{(\sum_i w_i^*)^2}, \tag{20}$$

where

$$w_i = \frac{n_{T_i} n_{C_i}}{n_{T_i} + n_{C_i} - 1}. \tag{21}$$

The confidence interval estimate of δ is obtained by:

$$d_{MH} \pm |z_{\alpha/2}|\, S_{MH}. \tag{22}$$

Method

The simulation study was programmed in GAUSS (Aptech Systems, 1992). Two binomially distributed populations were defined, $B(n_T, \pi_T)$ and $B(n_C, \pi_C)$, where π_T and π_C were the treated and control proportions, respectively. From these populations, pairs of independent random samples were generated with n_T and n_C as sample sizes. The simulated studies were accomplished assuming a treatment and a control group and a dichotomous outcome variable. Thus, a 2×2 table represented the data of each simulated study.

Each 2×2 table simulated the data from a primary research, where the sample risk difference ($d_i = p_{T_i} - p_{C_i}$) was computed. A set of k 2×2 independent tables simulating the data of a meta-analysis was run, yielding k risk

differences. All the 2×2 tables within the same meta-analysis estimated a common population risk difference, $\delta = \pi_T - \pi_C$.

The following factors were manipulated: (a) the average sample size of each meta-analysis, $\overline{N}(\overline{N} = \Sigma N_i/k_i$, and being $N_i = n_{T_i} + n_{C_i})$, with values 60, 100, and 160; (b) the number of studies, k, with values 10, 20, and 40; (c) the population risk difference, with values $\delta = \pi_T - \pi_C = 0$, 0.05, and 0.10; (d) the position in the range of the two population proportions, with the conditions central (0.5 vs. 0.5, 0.525 vs. 0.475, and 0.55 vs. 0.45) and extreme (0.1 vs. 0.1, 0.125 vs. 0.075, 0.15 vs. 0.05, 0.05 vs. 0.05, and 0.075 vs. 0.025); and (e) the relationship between the two population proportions and sample sizes, with the most extreme proportion assigned to the lowest sample size and vice versa (direct vs. indirect relationship).

Another factor taken into account was the ratio between the sample sizes of the two groups in each study. To study the influence of such a factor, meta-analyses were simulated that presented a constant imbalance between sample sizes through the k studies in the same meta-analysis, with the three manipulated conditions $n_T/n_C = 1$, $n_T/n_C = 2$, and $n_T/n_C = 4$. On the other hand, for a more realistic situation, meta-analyses were also simulated that involved variable ratios between sample sizes, with approximately a third of the k studies showing one of the three conditions of imbalance mentioned above.

To generate the sample sizes, N, of the k 2×2 tables in a meta-analysis, some properties of the sample size distribution in 30 real meta-analyses in the behavioral sciences field were assessed. The meta-analyses were selected through a literature search on *PsycLIT* and *Current Contents: Social and Behavioral Sciences* from 1981 to 1994. To be included in our study of the sample size distribution, the meta-analysis had to report the sample sizes of each of the primary studies. The 30 selected meta-analyses were published in relevant journals in the behavioral sciences, such as *American Psychologist*, *Clinical Psychology Review*, *Journal of Applied Psychology*, *Journal of Consulting and Clinical Psychology*, *Journal of Educational Psychology*, *Journal of Personality and Social Psychology*, and *Psychological Bulletin*. The selected meta-analyses presented a number of studies, k, from 7 to 235, the mean and the median being 42 and 30 studies, respectively.

In particular, the Pearson skewness index of the distribution, computed for all the meta-analyses, gave a value of +1.464. In accordance with this value, three vectors of 10 Ns each were selected: [24, 24, 32, 32, 36, 36, 40, 40, 168, 168], [64, 64, 72, 72, 76, 76, 80, 80, 208, 208], and [124, 124, 132, 132, 136, 136, 140, 140, 268, 268], all with skewness +1.464 and averaging 60, 100, and 160, respectively. These were the sample size distributions for meta-analyses with 10 studies. To get distributions of 20 and 40 studies, each vector was replicated 2 and 4 times, respectively.

For each one of the 369 conditions defined, 10,000 replications were generated by Monte Carlo simulation. Thus, 3,690,000 meta-analyses were sim-

ulated. Throughout the 10,000 replications of each condition, the five procedures to estimate the average risk difference were applied, and their bias, variability, adjustment of confidence level, and statistical power were assessed.

The bias of each of the five estimators, d_{ML}, d_C, d_{MH}, d_U, and d_{CW}, was assessed through the 10,000 replications of the same meta-analysis as the mean difference between the 10,000 observed averages and the constant population effect size, δ. In addition, the variability of the procedures was assessed by the standard deviation of d_{ML}, d_C, d_{MH}, d_U, and d_{CW} estimators, respectively, in the 10,000 replications of the same meta-analysis.

To assess the empirical adjustment of the confidence intervals over the five estimators, the proportion of intervals that included the true parametric value, δ, in the 10,000 replications of the same meta-analysis was calculated. The nominal confidence level was $1 - \alpha = 0.95$ in all of the applications. Finally, to assess the statistical power of the five confidence intervals, the null hypothesis H_0: $\delta = 0$ was tested. In conditions where $\delta \neq 0$, the statistical power was estimated as the proportion of intervals that did not include the zero value in the 10,000 replications of the same meta-analysis.

Results and Discussion

Bias of the Estimators

As expected from a large sample theory, our results show that the five estimators of δ are practically unbiased. As an example, in Table 1 we present the bias of the estimators when the two population proportions (π_T and π_C) are extreme ($\pi_T = 0.150$ and $\pi_C = 0.050$) and more extreme ($\pi_T = 0.075$ and $\pi_C = 0.025$). It must be noted that the results in this table are based in meta-analyses in which different imbalances in sample sizes have been mixed. Nevertheless, the d_{CW} estimator shows a negative bias when the population proportions are in an extreme position, and the most extreme proportion is associated with the largest sample size. For example, with $k = 20$ studies, average sample size $\overline{N} = 60$, and $\pi_T = 0.075$ and $\pi_C = 0.0250$, the estimated bias is –0.0316. The dependence between the sample risk differences, d_i, and the estimated weights, w_i, produces such a negative bias. In these conditions, sample risk differences with different sign (plus or minus) to that of the population effect size receive a disproportionate weight, leading to negative bias.

On the other hand, in conditions where the most extreme proportion is associated with the largest sample size, the larger the ratio between sample sizes, the greater the negative bias in the d_{CW} estimator. Table 2 presents the bias of the five estimators as a function of the population proportions, π_T and π_C, the number of studies, k, average sample size, $\overline{N}$, and the ratio between

Table 1
Bias of the Estimators

		$\pi_T = 0.150$, $\pi_C = 0.050$, $\delta = 0.100$				$\pi_T = 0.075$, $\pi_C = 0.025$, $\delta = 0.050$			
k	$\overline{N}$	d_{ML}	d_C[a]	d_{CW}	d_U	d_{ML}	d_C[a]	d_{CW}	d_U
10	60*d*[b]	–0.0000	–0.0000	–0.0011	–0.0003	–0.0002	–0.0001	–0.0027	–0.0001
10	60*i*[c]	–0.0001	–0.0001	–0.0403	–0.0005	0.0000	0.0000	–0.0291	0.0002
10	100*d*	0.0000	–0.0000	0.0031	0.0000	0.0001	0.0001	–0.0017	0.0001
10	100*i*	–0.0001	–0.0001	–0.0216	–0.0000	0.0001	0.0002	–0.0248	0.0002
10	160*d*	–0.0001	–0.0000	0.0031	–0.0000	–0.0001	–0.0001	0.0008	–0.0001
10	160*i*	0.0000	–0.0000	–0.0102	–0.0001	–0.0001	–0.0001	–0.0156	–0.0001
20	60*d*	0.0001	0.0001	–0.0013	0.0000	–0.0001	–0.0001	–0.0029	–0.0002
20	60*i*	–0.0001	–0.0001	–0.0435	–0.0002	0.0000	0.0001	–0.0316	0.0001
20	100*d*	0.0001	0.0001	0.0033	0.0001	–0.0001	–0.0002	–0.0023	–0.0001
20	100*i*	0.0002	0.0002	–0.0244	0.0002	0.0001	0.0001	–0.0276	0.0000
20	160*d*	0.0000	0.0000	0.0033	–0.0000	–0.0001	–0.0001	0.0006	–0.0001
20	160*i*	0.0001	0.0001	–0.0109	0.0002	0.0001	0.0001	–0.0177	0.0001
40	60*d*	–0.0001	–0.0002	–0.0019	–0.0001	0.0001	0.0001	–0.0027	0.0001
40	60*i*	–0.0000	–0.0000	–0.0453	–0.0002	0.0000	0.0000	–0.0330	0.0002
40	100*d*	0.0000	–0.0000	0.0033	–0.0000	–0.0000	–0.0000	–0.0026	–0.0001
40	100*i*	–0.0000	–0.0000	–0.0264	–0.0000	0.0001	0.0001	–0.0290	0.0001
40	160*d*	–0.0001	–0.0000	0.0034	–0.0000	–0.0001	–0.0001	0.0004	–0.0001
40	160*i*	0.0001	0.0000	–0.0117	0.0000	0.0000	0.0000	–0.0191	0.0000

Note. The ratio between sample sizes is variable through the studies in a meta-analysis. π_T and π_C = population proportions, k = number of studies, $\overline{N}$ = average sample size, and $\delta = \pi_T - \pi_C$ = population risk difference.
a. The d_c and d_{MH} estimators show the same value and, as a consequence, both estimators yield the same bias.
b. *d* denotes a direct relationship where the most extreme proportion is associated with the lowest sample size.
c. *i* denotes an inverse relationship where the most extreme proportion is associated with the largest sample size.

sample sizes, n_T/n_C. For example, for $k = 20$, $\overline{N} = 100$, and most extreme population proportions ($\pi_T = 0.075$ and $\pi_C = 0.025$), Table 2 shows that the bias for d_{CW} estimator is –0.0110, –0.0263, and –0.0503, for n_T/n_C = 1, 2, and 4, respectively. In this table, the ratio between sample sizes remains constant through the k studies in each of the meta-analyses. In these conditions, the estimates of d_{ML}, d_C, and d_{MH} coincide, obtaining the same bias values.

Variability of the Estimators

Tables 3 to 5 present the estimated standard deviations and relative efficiency of the five estimators as a function of the number of studies, k, average sample size, $\overline{N}$, population proportions, π_T and π_C, population risk difference, δ, and the sample size-population proportions relationship. Note that these tables are based on meta-analyses in which we have mixed different ratios in sample sizes. Consistent with large sample theory, as the sample size and the

Table 2
Bias of the Estimators

		$\pi_T = 0.075$, $\pi_C = 0.025$, $\delta = 0.050$								
		$k = 10$			$k = 20$			$k = 40$		
$\overline{N}$	n_T/n_C	d_{ML}[a]	d_{CW}	d_U	d_{ML}[a]	d_{CW}	d_U	d_{ML}[a]	d_{CW}	d_U
60	1	0.0002	–0.0137	0.0003	–0.0001	–0.0150	–0.0003	–0.0001	–0.0155	0.0001
60	2*d*[b]	0.0004	–0.0008	0.0003	–0.0000	–0.0014	–0.0000	–0.0000	–0.0017	0.0000
60	2*i*[c]	–0.0000	–0.0292	0.0002	–0.0001	–0.0309	–0.0002	–0.0000	–0.0317	0.0000
60	4*d*	0.0000	0.0088	0.0001	0.0001	0.0088	0.0001	0.0002	0.0088	0.0001
60	4*i*	0.0002	–0.0496	0.0002	0.0000	–0.0521	–0.0002	–0.0001	–0.0536	–0.0002
100	1	0.0000	–0.0098	0.0000	–0.0000	–0.0110	0.0000	0.0000	–0.0116	0.0001
100	2*d*	0.0000	–0.0009	0.0000	–0.0000	–0.0014	–0.0000	–0.0000	–0.0016	–0.0000
100	2*i*	0.0003	–0.0238	0.0003	–0.0001	–0.0263	–0.0001	0.0001	–0.0273	0.0001
100	4*d*	–0.0001	0.0065	–0.0002	–0.0000	0.0061	0.0000	–0.0000	0.0058	–0.0000
100	4*i*	–0.0000	–0.0472	0.0000	–0.0002	–0.0503	–0.0001	–0.0001	–0.0515	–0.0001
160	1	0.0000	–0.0053	0.0000	0.0000	–0.0059	0.0000	–0.0000	–0.0062	–0.0000
160	2*d*	0.0002	0.0009	0.0002	–0.0001	0.0005	–0.0001	–0.0000	0.0004	0.0000
160	2*i*	–0.0002	–0.0146	–0.0002	–0.0000	–0.0160	–0.0001	–0.0000	–0.0168	–0.0000
160	4*d*	0.0001	0.0079	0.0001	–0.0001	0.0078	–0.0000	–0.0001	0.0077	–0.0001
160	4*i*	–0.0000	–0.0336	–0.0000	0.0001	–0.0381	0.0001	–0.0000	–0.0405	–0.0000
		$\pi_T = 0.150$, $\pi_C = 0.050$, $\delta = 0.100$								
		$k = 10$			$k = 20$			$k = 40$		
$\overline{N}$	n_T/n_C	d_{ML}[a]	d_{CW}	d_U	d_{ML}[a]	d_{CW}	d_U	d_{ML}[a]	d_{CW}	d_U
60	1	0.0002	–0.0142	0.0005	0.0002	–0.0154	0.0000	0.0001	–0.0165	0.0001
60	2*d*[b]	0.0000	0.0006	0.0000	0.0001	0.0001	0.0000	0.0002	–0.0002	0.0003
60	2*i*[c]	–0.0001	–0.0383	0.0002	0.0001	–0.0412	–0.0000	–0.0001	–0.0428	–0.0002
60	4*d*	0.0000	0.0130	0.0001	0.0000	0.0129	0.0000	–0.0001	0.0127	–0.0003
60	4*i*	0.0004	–0.0801	0.0009	0.0001	–0.0851	–0.0000	–0.0003	–0.0878	–0.0004
100	1	–0.0002	–0.0066	–0.0000	–0.0001	–0.0072	–0.0002	0.0000	–0.0073	–0.0001
100	2*d*	–0.0000	0.0030	0.0002	0.0001	0.0031	0.0001	–0.0002	0.0027	–0.0002
100	2*i*	0.0002	–0.0191	0.0003	–0.0001	–0.0215	–0.0001	–0.0000	–0.0226	0.0000
100	4*d*	–0.0001	0.0154	–0.0001	–0.0000	0.0157	0.0001	0.0001	0.0159	0.0002
100	4*i*	–0.0002	–0.0510	–0.0003	–0.0000	–0.0588	–0.0001	–0.0002	–0.0638	–0.0002
160	1	–0.0001	–0.0036	–0.0001	0.0001	–0.0036	0.0001	0.0000	–0.0038	0.0000
160	2*d*	0.0000	0.0027	–0.0000	–0.0001	0.0026	–0.0001	–0.0002	0.0026	–0.0002
160	2*i*	0.0002	–0.0098	0.0002	–0.0000	–0.0108	–0.0000	0.0001	–0.0112	0.0001
160	4*d*	–0.0001	0.0123	–0.0001	0.0001	0.0131	0.0001	0.0001	0.0133	0.0001
160	4*i*	–0.0002	–0.0246	–0.0003	0.0001	–0.0273	0.0001	0.0002	–0.0295	0.0002

Note. The ratio between sample sizes is constant in the k studies of the same meta-analysis. π_T and π_C = population proportions, k = number of studies, $\overline{N}$ = average sample size, n_T/n_C = ratio between sample sizes, and $\delta = \pi_T - \pi_C$ = population risk difference.

a. In these conditions, the d_{ML}, d_c, and d_{MH} estimators show the same value and, as a consequence, the three estimators yield the same bias.

b. *d* denotes a direct relationship where the most extreme proportion is associated with the lowest sample size.

c. *i* denotes an inverse relationship where the most extreme proportion is associated with the largest sample size.

Table 3
Standard Deviations and Efficiency (in parentheses) of the Estimators

		$\pi_T = 0.500$, $\pi_C = 0.500$, $\delta = 0.000$				$\pi_T = 0.100$, $\pi_C = 0.100$, $\delta = 0.000$				$\pi_T = 0.050$, $\pi_C = 0.050$, $\delta = 0.000$			
k	$\overline{N}$	d_{ML}	d_C^a	d_{CW}	d_U	d_{ML}	d_C^a	d_{CW}	d_U	d_{ML}	d_C^a	d_{CW}	d_U
10	60	0.0437	0.0450	0.0515	0.0560	0.0264	0.0273	0.0243	0.0337	0.0194	0.0199	0.0151	0.0244
			(94.25)	(71.84)	(60.84)		(93.24)	(117.9)	(61.29)		(94.54)	(164.6)	(63.07)
10	100	0.0338	0.0348	0.0362	0.0380	0.0200	0.0206	0.0217	0.0226	0.0147	0.0152	0.0132	0.0166
			(94.20)	(87.11)	(79.01)		(94.11)	(85.06)	(78.72)		(93.41)	(124.3)	(78.14)
10	160	0.0263	0.0272	0.0278	0.0283	0.0157	0.0163	0.0175	0.0170	0.0115	0.0119	0.0120	0.0124
			(93.93)	(89.88)	(86.58)		(93.58)	(80.49)	(85.92)		(93.96)	(91.84)	(86.38)
20	60	0.0308	0.0318	0.0368	0.0394	0.0186	0.0192	0.0168	0.0239	0.0133	0.0137	0.0104	0.0169
			(94.30)	(70.41)	(61.18)		(94.33)	(122.8)	(60.58)		(93.43)	(162.2)	(61.44)
20	100	0.0236	0.0245	0.0255	0.0268	0.0144	0.0149	0.0156	0.0163	0.0103	0.0106	0.0089	0.0116
			(92.91)	(85.31)	(77.64)		(93.03)	(84.40)	(78.01)		(93.95)	(133.3)	(78.05)
20	160	0.0186	0.0192	0.0196	0.0201	0.0113	0.0116	0.0127	0.0121	0.0082	0.0085	0.0085	0.0088
			(93.43)	(89.32)	(85.28)		(93.76)	(78.66)	(86.74)		(93.14)	(92.57)	(86.48)
40	60	0.0219	0.0226	0.0263	0.0279	0.0130	0.0135	0.0117	0.0167	0.0095	0.0098	0.0074	0.0122
			(94.16)	(69.09)	(61.42)		(93.35)	(122.5)	(60.50)		(93.83)	(167.3)	(61.00)
40	100	0.0167	0.0174	0.0182	0.0190	0.0101	0.0104	0.0109	0.0113	0.0073	0.0075	0.0062	0.0082
			(92.54)	(84.85)	(77.76)		(93.82)	(86.05)	(79.84)		(93.98)	(139.7)	(80.40)
40	160	0.0132	0.0136	0.0139	0.0142	0.0080	0.0082	0.0090	0.0086	0.0057	0.0059	0.0059	0.0062
			(93.73)	(89.58)	(85.93)		(93.74)	(78.58)	(86.25)		(92.63)	(93.95)	(85.04)

Note. The ratio between sample sizes is variable through the studies in a meta-analysis. π_T and π_C = population proportions, k = number of studies, $\overline{N}$ = average sample size, and $\delta = \pi_T - \pi_C$ = population risk difference.
a. The d_c and d_{MH} estimators show the same value and, as a consequence, both estimators yield the same standard deviation.

Table 4
Standard Deviations and Efficiency (in parentheses) of the Estimators

		$\pi_T = 0.525$, $\pi_C = 0.475$, $\delta = 0.050$				$\pi_T = 0.55$, $\pi_C = 0.45$, $\delta = 0.100$			
k	$\overline{N}$	d_{ML}	d_C^a	d_{CW}	d_U	d_{ML}	d_C^a	d_{CW}	d_U
10	60	0.0437	0.0451	0.0517	0.0561	0.0434	0.0446	0.0510	0.0548
			(93.88)	(71.52)	(60.78)		(94.60)	(72.47)	(62.78)
10	100	0.0332	0.0345	0.0360	0.0378	0.0333	0.0344	0.0358	0.0376
			(92.53)	(85.15)	(77.09)		(93.38)	(86.48)	(78.16)
10	160	0.0262	0.0272	0.0278	0.0284	0.0261	0.0270	0.0277	0.0283
			(93.17)	(89.26)	(84.95)		(93.51)	(89.36)	(85.43)
20	60	0.0311	0.0321	0.0370	0.0393	0.0309	0.0320	0.0374	0.0397
			(93.92)	(70.48)	(62.59)		(93.63)	(68.38)	(60.65)
20	100	0.0236	0.0244	0.0254	0.0266	0.0235	0.0242	0.0253	0.0265
			(93.94)	(86.23)	(78.60)		(93.71)	(85.98)	(78.25)
20	160	0.0188	0.0193	0.0198	0.0202	0.0186	0.0191	0.0196	0.0199
			(94.15)	(89.96)	(85.95)		(94.36)	(90.16)	(87.29)
40	60	0.0217	0.0225	0.0262	0.0279	0.0219	0.0226	0.0265	0.0281
			(93.36)	(68.69)	(60.82)		(93.85)	(67.99)	(60.83)
40	100	0.0169	0.0175	0.0182	0.0191	0.0167	0.0172	0.0180	0.0188
			(93.60)	(86.29)	(78.98)		(93.98)	(86.15)	(78.48)
40	160	0.0131	0.0136	0.0139	0.0142	0.0132	0.0137	0.0140	0.0143
			(92.56)	(88.31)	(85.16)		(92.99)	(88.66)	(85.12)

Note. The ratio between sample sizes is variable through the studies in a meta-analysis. π_T and π_C = population proportions, k = number of studies, $\overline{N}$ = average sample size, and $\delta = \pi_T - \pi_C$ = population risk difference.
a. The d_c and d_{MH} estimators show the same value and, as a consequence, both estimators yield the same standard deviation.

number of studies increase the estimators' variability decreases, achieving a larger efficiency. Furthermore, the more extreme the position of the population proportions, the lower the standard deviation. On the other hand, when the most extreme proportion is associated with the largest sample size, the standard deviations are larger than if the situation were reversed. For example, with $k = 20$, $\overline{N} = 60$, and $\pi_T = 0.15$ and $\pi_C = 0.050$, the d_{ML} estimator obtains standard deviations of 0.0168 and 0.0199, for direct and inverse relationships, respectively (see Table 5).

As the ratio between sample sizes increases, the estimators' variability also increases. Table 6 shows this trend in conditions where the population proportions are most extreme ($\pi_T = 0.075$ and $\pi_C = 0.025$). Note that in this table the ratio between sample sizes remains constant through the k studies in each of the meta-analyses. In these conditions, the estimates of d_{ML}, d_C, and d_{MH} coincide, obtaining the same standard deviations. For example, for $\overline{N} = 100$, $k = 20$, and an inverse sample size-population proportion relationship, d_{ML}, d_C, and d_{MH} estimators had standard deviations of 0.0111 and 0.0138, for $n_T/n_C = 2$ and 4, respectively. However, the d_{CW} estimator does not show a clear trend as a function of the ratio of sample sizes.

Table 5
Standard Deviations and Efficiency (in parentheses) of the Estimators

		$\pi_T = 0.125, \pi_C = 0.075, \delta = 0.050$				$\pi_T = 0.150, \pi_C = 0.050, \delta = 0.100$				$\pi_T = 0.075, \pi_C = 0.0250, \delta = 0.050$			
k	$\overline{N}$	d_{ML}	d_C^{a}	d_{CW}	d_U	d_{ML}	d_C^{a}	d_{CW}	d_U	d_{ML}	d_C^{a}	d_{CW}	d_U
10	$60d^{b}$	0.0253	0.0262	0.0234	0.0328	0.0237	0.0247	0.0241	0.0308	0.0174	0.0182	0.0150	0.0228
			(92.91)	(116.8)	(59.46)		(91.56)	(96.21)	(59.27)		(91.51)	(136.2)	(58.37)
10	$60i^{c}$	0.0271	0.0279	0.0268	0.0343	0.0278	0.0284	0.0326	0.0352	0.0207	0.0211	0.0181	0.0258
			(94.86)	(102.1)	(62.70)		(95.74)	(72.89)	(62.35)		(95.65)	(130.4)	(64.14)
10	100d	0.0194	0.0201	0.0201	0.0221	0.0183	0.0192	0.0191	0.0208	0.0133	0.0140	0.0133	0.0153
			(92.63)	(92.58)	(77.25)		(91.15)	(91.52)	(77.14)		(90.21)	(100.4)	(75.86)
10	100i	0.0208	0.0215	0.0253	0.0234	0.0213	0.0219	0.0299	0.0240	0.0156	0.0159	0.0184	0.0174
			(93.80)	(67.48)	(79.03)		(94.53)	(50.93)	(79.04)		(95.14)	(71.78)	(79.64)
10	160d	0.0150	0.0157	0.0163	0.0164	0.0145	0.0152	0.0151	0.0156	0.0105	0.0110	0.0107	0.0113
			(91.02)	(84.81)	(84.03)		(91.77)	(91.83)	(86.61)		(91.55)	(95.53)	(85.75)
10	160i	0.0163	0.0169	0.0191	0.0177	0.0166	0.0170	0.0199	0.0179	0.0123	0.0126	0.0171	0.0133
			(93.30)	(72.84)	(84.36)		(95.23)	(70.22)	(86.69)		(94.29)	(51.22)	(85.28)
20	60d	0.0178	0.0184	0.0166	0.0231	0.0168	0.0174	0.0172	0.0217	0.0121	0.0127	0.0103	0.0159
			(93.48)	(115.0)	(59.53)		(92.26)	(94.51)	(59.47)		(91.95)	(139.7)	(58.47)
20	60i	0.0192	0.0196	0.0187	0.0244	0.0199	0.0203	0.0230	0.0251	0.0147	0.0151	0.0134	0.0186
			(95.59)	(104.9)	(61.51)		(95.83)	(74.80)	(62.90)		(95.02)	(121.2)	(62.73)
20	100d	0.0136	0.0142	0.0143	0.0154	0.0129	0.0135	0.0136	0.0147	0.0095	0.0099	0.0096	0.0108
			(91.91)	(90.52)	(77.80)		(91.52)	(89.58)	(77.20)		(91.81)	(97.31)	(76.54)

20	100*i*	0.0148	0.0152	0.0184	0.0167	0.0152	0.0156	0.0230	0.0170	0.0112	0.0115	0.0129	0.0125
			(94.19)	(64.74)	(78.36)		(94.52)	(43.45)	(79.23)		(95.18)	(75.68)	(80.05)
20	160*d*	0.0107	0.0112	0.0115	0.0116	0.0103	0.0107	0.0106	0.0111	0.0074	0.0078	0.0077	0.0080
			(92.35)	(86.77)	(85.38)		(91.91)	(93.54)	(85.83)		(91.36)	(91.62)	(85.80)
20	160*i*	0.0115	0.0118	0.0138	0.0124	0.0117	0.0120	0.0148	0.0126	0.0087	0.0089	0.0131	0.0094
			(94.49)	(69.96)	(85.91)		(95.44)	(62.90)	(86.38)		(95.04)	(44.12)	(85.88)
40	60*d*	0.0126	0.0131	0.0116	0.0163	0.0118	0.0123	0.0123	0.0155	0.0086	0.0090	0.0073	0.0112
			(92.99)	(117.9)	(60.50)		(91.78)	(92.98)	(58.28)		(91.46)	(139.5)	(58.48)
40	60*i*	0.0136	0.0139	0.0131	0.0172	0.0138	0.0142	0.0160	0.0173	0.0102	0.0104	0.0097	0.0129
			(95.33)	(107.4)	(62.68)		(95.23)	(75.00)	(63.81)		(95.97)	(111.0)	(62.99)
40	100*d*	0.0097	0.0101	0.0101	0.0110	0.0090	0.0095	0.0096	0.0103	0.0066	0.0069	0.0068	0.0075
			(92.99)	(92.08)	(78.36)		(91.43)	(89.17)	(76.59)		(92.54)	(94.15)	(78.23)
40	100*i*	0.0104	0.0107	0.0130	0.0117	0.0107	0.0110	0.0169	0.0120	0.0078	0.0080	0.0090	0.0087
			(94.20)	(64.63)	(79.45)		(94.09)	(39.57)	(79.39)		(95.85)	(75.76)	(80.29)
40	160*d*	0.0075	0.0079	0.0082	0.0081	0.0071	0.0075	0.0075	0.0077	0.0053	0.0056	0.0056	0.0058
			(92.15)	(84.87)	(85.68)		(90.65)	(91.25)	(84.90)		(90.18)	(88.18)	(84.48)
40	160*i*	0.0082	0.0084	0.0098	0.0088	0.0084	0.0086	0.0112	0.0090	0.0061	0.0062	0.0094	0.0066
			(94.14)	(69.16)	(86.65)		(94.66)	(55.56)	(85.74)		(95.59)	(41.48)	(86.09)

Note. The ratio between sample sizes is variable through the studies in a meta-analysis. π_T and π_C = population proportions, k = number of studies, $\overline{N}$ = average sample size, and $\delta = \pi_T - \pi_C$ = population risk difference.

a. The d_c and d_{MH} estimators show the same value and, as a consequence, both estimators yield the same bias.

b. *d* denotes a direct relationship where the most extreme proportion is associated with the lowest sample size.

c. *i* denotes an inverse relationship where the most extreme proportion is associated with the largest sample size.

Table 6
Standard Deviations and Efficiency (in parentheses) of the Estimators

		$\pi_T = 0.075$, $\pi_C = 0.025$, $\delta = 0.050$								
		$k = 10$			$k = 20$			$k = 40$		
$\overline{N}$	n_T/n_C	d_{ML}[a]	d_{CW}	d_U	d_{ML}[a]	d_{CW}	d_U	d_{ML}[a]	d_{CW}	d_U
60	1	0.0179	0.0156	0.0225	0.0125	0.0106	0.0157	0.0089	0.0076	0.0112
			(132.4)	(63.52)		(138.7)	(63.49)		(136.3)	(62.61)
60	2*d*[b]	0.0173	0.0141	0.0217	0.0122	0.0097	0.0152	0.0086	0.0069	0.0108
			(151.4)	(63.80)		(156.0)	(63.91)		(154.2)	(62.89)
60	2*i*[c]	0.0201	0.0164	0.0254	0.0143	0.0114	0.0180	0.0101	0.0081	0.0127
			(149.5)	(62.12)		(157.6)	(63.42)		(155.1)	(62.38)
60	4*d*	0.0185	0.0124	0.0234	0.0132	0.0087	0.0168	0.0093	0.0063	0.0117
			(223.3)	(62.10)		(227.5)	(62.02)		(221.8)	(63.57)
60	4*i*	0.0253	0.0173	0.0315	0.0176	0.0115	0.0224	0.0127	0.0080	0.0160
			(214.2)	(64.47)		(235.2)	(62.05)		(250.3)	(62.86)
100	1	0.0137	0.0142	0.0150	0.0097	0.0101	0.0106	0.0069	0.0073	0.0075
			(93.91)	(83.96)		(91.22)	(83.56)		(87.79)	(82.60)
100	2*d*	0.0131	0.0121	0.0144	0.0093	0.0085	0.0103	0.0066	0.0061	0.0072
			(118.7)	(83.03)		(119.1)	(81.91)		(115.7)	(83.08)
100	2*i*	0.0154	0.0174	0.0169	0.0111	0.0122	0.0122	0.0077	0.0082	0.0085
			(78.59)	(83.15)		(83.30)	(82.12)		(88.30)	(82.78)
100	4*d*	0.0145	0.0118	0.0160	0.0102	0.0083	0.0112	0.0073	0.0060	0.0080
			(150.8)	(81.88)		(149.0)	(82.54)		(150.8)	(84.01)
100	4*i*	0.0195	0.0172	0.0215	0.0138	0.0102	0.0151	0.0097	0.0064	0.0106
			(128.5)	(82.53)		(182.3)	(83.24)		(230.0)	(82.75)
160	1	0.0108	0.0112	0.0112	0.0077	0.0082	0.0080	0.0054	0.0058	0.0056
			(93.25)	(93.70)		(88.22)	(91.40)		(85.19)	(92.05)
160	2*d*	0.0105	0.0101	0.0109	0.0076	0.0072	0.0079	0.0053	0.0051	0.0055
			(109.6)	(93.01)		(109.8)	(92.37)		(106.5)	(91.70)
160	2*i*	0.0125	0.0154	0.0130	0.0087	0.0114	0.0090	0.0061	0.0083	0.0064
			(66.08)	(92.42)		(58.35)	(92.53)		(54.10)	(91.59)
160	4*d*	0.0113	0.0098	0.0118	0.0081	0.0071	0.0085	0.0057	0.0050	0.0059
			(132.3)	(91.62)		(130.0)	(91.97)		(126.2)	(92.44)
160	4*i*	0.0154	0.0216	0.0159	0.0108	0.0144	0.0113	0.0077	0.0095	0.0081
			(50.77)	(92.80)		(56.30)	(91.54)		(67.03)	(91.39)

Note. The ratio between sample sizes is constant in the k studies of the same meta-analysis. π_T and π_C = population proportions, k = number of studies, $\overline{N}$ = average sample size, n_T/n_C = ratio between sample sizes, and $\delta = \pi_T - \pi_C$ = population risk difference.

a. In these conditions, the d_{ML}, d_C, and d_{MH} estimators show the same value and, as a consequence, the three estimators yield the same standard deviation.

b. *d* denotes a direct relationship where the most extreme proportion is associated with the lowest sample size.

c. *i* denotes an inverse relationship where the most extreme proportion is associated with the largest sample size.

To compare the variability of the five estimators of a common risk difference, the relative efficiency was computed by taking the variance of the d_{ML} estimator divided by the variance of each of the remaining estimators and multiplied by 100. Tables 3 to 5 present the relative efficiencies of d_C, d_{MH},

Table 7
Empirical Confidence Level for the Five Procedures

		$\pi_T = 0.500$, $\pi_C = 0.500$, $\delta = 0.000$					$\pi_T = 0.100$, $\pi_C = 0.100$, $\delta = 0.000$					$\pi_T = 0.050$, $\pi_C = 0.050$, $\delta = 0.000$				
k	$\overline{N}$	d_{ML}	d_C	d_{MH}	d_{CW}	d_U	d_{ML}	d_C	d_{MH}	d_{CW}	d_U	d_{ML}	d_C	d_{MH}	d_{CW}	d_U
10	60	.9465	.9489	.9501	.9039	.9533	.9462	.9469	.9494	.8204	.9473	.9445	.9470	.9493	.8048	.9502
10	100	.9479	.9480	.9493	.9363	.9539	.9482	.9508	.9518	.8299	.9528	.9476	.9496	.9505	.7808	.9493
10	160	.9499	.9513	.9518	.9446	.9530	.9503	.9510	.9514	.8767	.9538	.9501	.9522	.9535	.7988	.9535
20	60	.9508	.9488	.9504	.9012	.9523	.9489	.9510	.9522	.7104	.9492	.9507	.9499	.9528	.6816	.9522
20	100	.9506	.9514	.9523	.9356	.9506	.9471	.9468	.9476	.7383	.9493	.9506	.9500	.9513	.6403	.9517
20	160	.9493	.9502	.9509	.9427	.9491	.9472	.9466	.9475	.8307	.9487	.9472	.9466	.9468	.6900	.9520
40	60	.9470	.9461	.9485	.8966	.9505	.9529	.9495	.9518	.5003	.9508	.9494	.9502	.9527	.4479	.9458
40	100	.9489	.9462	.9473	.9329	.9482	.9497	.9484	.9496	.5964	.9498	.9480	.9455	.9469	.3929	.9527
40	160	.9500	.9501	.9505	.9434	.9507	.9506	.9488	.9497	.7622	.9492	.9518	.9527	.9531	.5148	.9542

Note. The ratio between sample sizes is variable through the studies in a meta-analysis. π_T and π_C = population proportions, k = number of studies, $\overline{N}$ = average sample size, and $\delta = \pi_T - \pi_C$ = population risk difference.

Table 8
Empirical Confidence Level and Statistical Power (in parentheses) of the Five Procedures

		$\pi_T = 0.525$, $\pi_C = 0.475$, $\delta = 0.050$					$\pi_T = 0.550$, $\pi_C = 0.450$, $\delta = 0.100$				
k	$\overline{N}$	d_{ML}	d_C	d_{MH}	d_{CW}	d_U	d_{ML}	d_C	d_{MH}	d_{CW}	d_U
10	60	.9499	.9491	.9521	.9018	.9499	.9491	.9508	.9525	.8994	.9584
		(.2134)	(.2017)	(.1964)	(.2612)	(.1275)	(.6265)	(.6048)	(.5970)	(.6510)	(.3657)
10	100	.9500	.9476	.9486	.9352	.9505	.9505	.9480	.9493	.9325	.9520
		(.3313)	(.3111)	(.3072)	(.3386)	(.2349)	(.8531)	(.8262)	(.8238)	(.8440)	(.6590)
10	160	.9500	.9494	.9505	.9404	.9490	.9483	.9501	.9508	.9404	.9511
		(.4774)	(.4568)	(.4547)	(.4751)	(.3612)	(.9667)	(.9594)	(.9587)	(.9617)	(.8720)
20	60	.9478	.9464	.9490	.8972	.9544	.9497	.9479	.9497	.8882	.9474
		(.3664)	(.3569)	(.3492)	(.4149)	(.2296)	(.8994)	(.8831)	(.8801)	(.8964)	(.6715)
20	100	.9504	.9501	.9513	.9350	.9523	.9493	.9515	.9537	.9354	.9515
		(.5671)	(.5423)	(.5368)	(.5693)	(.4320)	(.9884)	(.9845)	(.9839)	(.9858)	(.9423)
20	160	.9473	.9471	.9477	.9412	.9481	.9485	.9495	.9505	.9407	.9533
		(.7601)	(.7316)	(.7299)	(.7442)	(.6475)	(.9999)	(.9997)	(.9997)	(.9997)	(.9966)
40	60	.9533	.9483	.9504	.8943	.9497	.9479	.9474	.9501	.8798	.9482
		(.6360)	(.6167)	(.6090)	(.6597)	(.4182)	(.9969)	(.9941)	(.9939)	(.9928)	(.9328)
40	100	.9467	.9448	.9465	.9308	.9503	.9520	.9509	.9520	.9325	.9503
		(.8436)	(.8216)	(.8191)	(.8389)	(.7318)	(.9999)	(1.0)	(1.0)	(1.0)	(.9997)
40	160	.9504	.9500	.9505	.9422	.9528	.9500	.9488	.9496	.9377	.9470
		(.9687)	(.9589)	(.9585)	(.9611)	(.9308)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)

Note. The ratio between sample sizes is variable through the studies in a meta-analysis. π_T and π_C = population proportions, k = number of studies, $\overline{N}$ = average sample size, and $\delta = \pi_T - \pi_C$ = population risk difference.

d_{CW}, and d_U estimators. As a general rule, the d_{ML} estimator showed the largest efficiency, closely followed by the d_C and d_{MH} estimators, with d_U being the least efficient estimator. On the other hand, the d_{CW} estimator exhibited a more irregular pattern, at times surpassing the efficiency of the d_{ML} estimator.

The conditions where the d_{CW} estimator achieved the largest efficiency were with extreme population proportions and average sample size over $\overline{N} = 60$. For example, with $k = 20$, $\overline{N} = 60$, and $\pi_T = \pi_C = 0.05$, the relative efficiency of the d_{CW} estimator is 162.2% (see Table 3). Although it has the lowest variability, in these conditions the d_{CW} estimator is negatively biased, and, as we will see later, it does not adequately adjust the confidence level.

Adjustment of the Confidence Level

Tables 7 to 9 show the adjustment of the confidence level for the confidence intervals constructed from the five estimators, the nominal confidence level being $1 - \alpha = 0.95$. When the population proportions are in a central position, d_{ML}, d_C, d_{MH}, and d_U estimators conform to the nominal confidence level. However, the d_{CW} estimator shows an inferior adjustment (see Tables 7 and 8), especially with small sample sizes ($\overline{N} = 60$).

Table 9
Empirical Confidence Level and Statistical Power (in parentheses) for the Five Procedures

		$\pi_T = 0.125$, $\pi_C = 0.075$, $\delta = 0.050$					$\pi_T = 0.15$, $\pi_C = 0.05$, $\delta = 0.100$					$\pi_T = 0.075$, $\pi_C = 0.025$, $\delta = 0.050$				
k	$\overline{N}$	d_{ML}	d_C	d_{MH}	d_{CW}	d_U	d_{ML}	d_C	d_{MH}	d_{CW}	d_U	d_{ML}	d_C	d_{MH}	d_{CW}	d_U
10	60d[a]	.9454	.9635	.9652	.8981	.9515	.9467	.9755	.9767	.9056	.9526	.9473	.9758	.9765	.9408	.9566
		(.5195)	(.4359)	(.4278)	(.7619)	(.3208)	(.9803)	(.9649)	(.9633)	(.9923)	(.8166)	(.8025)	(.7049)	(.6987)	(.8912)	(.5449)
10	60i[b]	.9511	.9314	.9347	.6555	.9549	.9467	.9115	.9149	.5163	.9522	.9403	.9013	.9044	.5421	.9468
		(.4394)	(.4884)	(.4818)	(.1459)	(.2358)	(.9673)	(.9723)	(.9714)	(.6513)	(.7371)	(.7029)	(.7709)	(.7656)	(.1980)	(.3917)
10	100d	.9477	.9645	.9655	.8812	.9516	.9485	.9764	.9772	.9147	.9550	.9524	.9766	.9773	.8939	.9521
		(.7264)	(.6628)	(.6585)	(.8632)	(.5384)	(.9996)	(.9988)	(.9987)	(.9998)	(.9838)	(.9560)	(.9191)	(.9168)	(.9804)	(.8322)
10	100i	.9465	.9309	.9321	.7280	.9520	.9423	.9078	.9097	.7067	.9489	.9441	.9123	.9142	.4399	.9481
		(.6864)	(.6997)	(.6965)	(.4067)	(.4852)	(.9991)	(.9989)	(.9989)	(.9202)	(.9721)	(.9274)	(.9365)	(.9354)	(.4576)	(.7510)
10	160d	.9516	.9634	.9642	.9015	.9537	.9443	.9735	.9741	.9226	.9483	.9485	.9767	.9771	.9188	.9519
		(.8992)	(.8601)	(.8589)	(.9359)	(.7617)	(1.0)	(1.0)	(1.0)	(1.0)	(.9998)	(.9974)	(.9924)	(.9922)	(.9974)	(.9665)
10	160i	.9463	.9335	.9344	.8540	.9507	.9504	.9232	.9241	.8579	.9544	.9476	.9137	.9149	.6287	.9525
		(.8858)	(.8792)	(.8778)	(.7224)	(.7250)	(1.0)	(1.0)	(1.0)	(.9961)	(.9994)	(.9918)	(.9918)	(.9913)	(.7800)	(.9385)
20	60d	.9509	.9648	.9659	.8773	.9496	.9473	.9756	.9770	.9002	.9523	.9515	.9775	.9789	.9380	.9542
		(.7902)	(.7300)	(.7238)	(.9596)	(.5617)	(.9999)	(.9997)	(.9997)	(.9999)	(.9840)	(.9776)	(.9564)	(.9548)	(.9954)	(.8333)
20	60i	.9493	.9323	.9347	.4523	.9509	.9473	.9090	.9126	.3060	.9480	.9424	.8996	.9028	.2644	.9417
		(.7521)	(.7682)	(.7633)	(.1932)	(.4909)	(.9996)	(.9996)	(.9996)	(.8923)	(.9808)	(.9563)	(.9605)	(.9591)	(.3533)	(.7788)
20	100d	.9514	.9625	.9634	.8558	.9538	.9512	.9755	.9762	.9067	.9525	.9496	.9744	.9753	.8772	.9485
		(.9500)	(.9263)	(.9247)	(.9893)	(.8540)	(1.0)	(1.0)	(1.0)	(1.0)	(.9999)	(.9991)	(.9973)	(.9973)	(.9999)	(.9873)
20	100i	.9467	.9314	.9329	.6071	.9469	.9467	.9156	.9169	.5607	.9466	.9481	.9091	.9105	.2248	.9478
		(.9369)	(.9315)	(.9301)	(.5750)	(.8332)	(1.0)	(1.0)	(1.0)	(.9956)	(1.0)	(.9986)	(.9984)	(.9984)	(.6598)	(.9848)

(continued)

Table 9 Continued

		$\pi_T = 0.125$, $\pi_C = 0.075$, $\delta = 0.050$					$\pi_T = 0.15$, $\pi_C = 0.05$, $\delta = 0.100$					$\pi_T = 0.075$, $\pi_C = 0.025$, $\delta = 0.050$				
k	$\overline{N}$	d_{ML}	d_C	d_{MH}	d_{CW}	d_U	d_{ML}	d_C	d_{MH}	d_{CW}	d_U	d_{ML}	d_C	d_{MH}	d_{CW}	d_U
20	160*d*	.9498	.9627	.9630	.8730	.9493	.9465	.9743	.9747	.9193	.9502	.9512	.9745	.9748	.9105	.9541
		(.9966)	(.9929)	(.9929)	(.9989)	(.9802)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
20	160*i*	.9502	.9359	.9364	.8028	.9513	.9483	.9198	.9210	.8008	.9516	.9454	.9138	.9143	.4339	.9483
		(.9935)	(.9915)	(.9914)	(.9116)	(.9736)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(.9415)	(.9996)
40	60*d*	.9488	.9624	.9642	.8325	.9495	.9491	.9774	.9782	.8965	.9513	.9518	.9792	.9802	.9265	.9509
		(.9752)	(.9586)	(.9571)	(.9996)	(.8441)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(.9995)	(.9993)	(.9993)	(1.0)	(.9870)
40	60*i*	.9529	.9338	.9367	.1934	.9515	.9522	.9124	.9151	.1033	.9535	.9480	.9047	.9081	.0539	.9473
		(.9643)	(.9654)	(.9641)	(.3019)	(.8175)	(1.0)	(1.0)	(1.0)	(.9904)	(1.0)	(1.0)	(.9997)	(.9996)	(.5548)	(.9855)
40	100*d*	.9500	.9628	.9640	.7793	.9494	.9497	.9763	.9771	.9009	.9485	.9505	.9749	.9756	.8599	.9524
		(.9988)	(.9984)	(.9984)	(.9999)	(.9919)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
40	100*i*	.9472	.9298	.9309	.4292	.9471	.9470	.9147	.9166	.3584	.9487	.9496	.9115	.9129	.0519	.9476
		(.9987)	(.9987)	(.9986)	(.8072)	(.9904)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(.8842)	(1.0)
40	160*d*	.9524	.9637	.9644	.8288	.9516	.9513	.9760	.9769	.9061	.9524	.9525	.9738	.9747	.9052	.9495
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
40	160*i*	.9468	.9319	.9327	.7070	.9493	.9490	.9187	.9202	.6790	.9474	.9515	.9187	.9196	.2047	.9493
		(1.0)	(1.0)	(1.0)	(.9932)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(.9981)	(1.0)

Note. The ratio between sample sizes is variable through the studies in a meta-analysis. π_T and π_C = population proportions, k = number of studies, $\overline{N}$ = average sample size, and $\delta = \pi_T - \pi_C$ = population risk difference.

a. *d* denotes a direct relationship where the most extreme proportion is associated with the lowest sample size.

b. *i* denotes an inverse relationship where the most extreme proportion is associated with the largest sample size.

When the population proportions are in an extreme position, the performance of the estimators is more variable (see Tables 7 and 9). First, d_{ML} and d_U estimators maintain their adequate adjustment of the confidence level. Second, the d_C and d_{MH} estimators also conform to the nominal confidence level when $\delta = 0$ (see Table 7). However, with $\delta \neq 0$, d_C and d_{MH} show a slightly irregular performance: When the sample size-population proportions relationship is inverse, these estimators underestimate the nominal confidence level, whereas with a direct relationship, the empirical values overestimate the nominal level. For example, with $k = 20$, $\overline{N} = 100$, $\delta = 0.10$, and $\pi_T = 0.15$ and $\pi_C = 0.05$, d_C and d_{MH} obtain empirical confidence levels of 0.9755 and 0.9762, respectively, for a direct relationship, and levels of 0.9156 and 0.9169 for an inverse relationship (see Table 9).

The most irregular performance was that of the d_{CW} estimator. Its empirical values systematically underestimated the nominal confidence level, improving the adjustment as the average sample size increased. A counterintuitive result was the deterioration of the adjustment as the number of studies increased (see Table 7). For example, with $\overline{N} = 100$ and $\pi_T = \pi_C = 0.10$, the empirical confidence values of d_{CW} were 0.8299, 0.7383, and 0.5964, for $k = 10$, 20, and 40, respectively. On the other hand, the underestimation of the confidence level in the d_{CW} estimator was more pronounced in an inverse than in a direct relationship between sample sizes and population proportions (see Table 9). For example, with $k = 20$, $\overline{N} = 100$, and $\pi_T = 0.075$ and $\pi_C = 0.025$, the empirical confidence values were 0.8772 and 0.2248, for a direct and an inverse relationship, respectively.

To examine the influence of the ratio between the treatment and control sample sizes on the adjustment of the confidence level, a part of our simulations held the ratio between n_T and n_C constant in all of the studies of each meta-analysis. As an example, Table 10 presents the empirical confidence levels of the five estimators for most extreme population proportions ($\pi_T = 0.075$ and $\pi_C = 0.025$). Neither the d_{ML} nor the d_U estimators were affected by the n_T/n_C ratio, the two procedures conforming to the nominal confidence level in all conditions. On the contrary, the d_C and d_{MH} estimators presented an interaction between the n_T/n_C ratio and the sample size-population proportion relationship. In particular, the unadjustment to the nominal confidence level increased as the ratio between sample sizes also increased, this unadjustment consisting of an underestimation in the inverse relationship and of an overestimation in the direct relationship. For example, for $k = 20$, $\overline{N} = 100$ and a direct relationship, the d_C estimator showed values of .9813 and .9909, for $n_T/n_C = 2$ and 4, respectively, whereas with an inverse relationship the corresponding values were .9017 and .8510.

On the other hand, the unadjustment of the d_{CW} estimator was larger as the ratio between sample sizes increased. For example, for $k = 20$, $\overline{N} = 100$ and an

Table 10
Empirical Confidence Level and Statistical Power (in parentheses) for the Five Procedures

		$\pi_T = 0.075, \pi_C = 0.025, \delta = 0.050$														
		$k = 10$					$k = 20$					$k = 40$				
$\overline{N}$	n_T/n_C	d_{ML}	d_C	d_{MH}	d_{CW}	d_U	d_{ML}	d_C	d_{MH}	d_{CW}	d_U	d_{ML}	d_C	d_{MH}	d_{CW}	d_U
60	1	.9415	.9410	.9446	.8191	.9481	.9492	.9493	.9504	.6770	.9515	.9498	.9492	.9514	.4415	.9452
		(.8259)	(.8264)	(.8249)	(.6724)	(.5425)	(.9836)	(.9847)	(.9836)	(.9330)	(.8686)	(1.0)	(1.0)	(1.0)	(.9990)	(.9938)
60	2*d*[a]	.9464	.9782	.9792	.9600	.9561	.9495	.9773	.9783	.9603	.9525	.9489	.9773	.9789	.9534	.9515
		(.8197)	(.7569)	(.7552)	(.9289)	(.5834)	(.9777)	(.9699)	(.9671)	(.9985)	(.8571)	(.9999)	(.9999)	(.9999)	(1.0)	(.9909)
60	2*i*[b]	.9453	.9029	.9098	.5204	.9471	.9489	.8993	.9041	.2363	.9426	.9491	.9039	.9071	.0381	.9488
		(.7239)	(.8059)	(.7988)	(.1916)	(.4051)	(.9623)	(.9754)	(.9744)	(.3600)	(.7860)	(.9998)	(.9998)	(.9998)	(.6464)	(.9869)
60	4*d*	.9440	.9905	.9910	.9274	.9467	.9440	.9902	.9907	.8831	.9448	.9446	.9899	.9907	.7731	.9472
		(.7372)	(.5459)	(.5425)	(.9915)	(.5389)	(.9407)	(.8888)	(.8866)	(.9999)	(.7789)	(.9979)	(.9960)	(.9957)	(1.0)	(.9645)
60	4*i*	.9289	.8434	.8453	.2053	.9302	.9421	.8473	.8539	.0317	.9379	.9436	.8491	.8529	.0004	.9386
		(.5110)	(.7200)	(.7061)	(.0639)	(.2064)	(.8513)	(.9241)	(.9238)	(.0688)	(.5679)	(.9909)	(.9956)	(.9954)	(.0969)	(.9171)
100	1	.9479	.9493	.9506	.7848	.9501	.9471	.9471	.9486	.6741	.9486	.9505	.9507	.9518	.4871	.9502
		(.9614)	(.9614)	(.9612)	(.9099)	(.8519)	(.9999)	(.9999)	(.9999)	(.9964)	(.9956)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
100	2*d*	.9532	.9813	.9816	.9325	.9535	.9525	.9813	.9815	.9271	.9480	.9534	.9796	.9803	.9148	.9549
		(.9582)	(.9387)	(.9384)	(.9896)	(.8538)	(.9991)	(.9991)	(.9991)	(1.0)	(.9929)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)

100	2*i*	.9492	.9088	.9109	.4659	.9489	.9449	.9017	.9047	.2425	.9471	.9505	.9057	.9071	.0577	.9510
		(.9272)	(.9517)	(.9509)	(.4811)	(.7708)	(.9988)	(.9994)	(.9994)	(.7106)	(.9863)	(1.0)	(1.0)	(1.0)	(.9351)	(1.0)
100	4*d*	.9400	.9907	.9907	.8883	.9469	.9470	.9909	.9915	.8462	.9480	.9467	.9907	.9914	.7751	.9459
		(.8952)	(.8067)	(.8041)	(.9993)	(.7702)	(.9942)	(.9871)	(.9866)	(1.0)	(.9722)	(1.0)	(1.0)	(1.0)	(1.0)	(.9994)
100	4*i*	.9375	.8483	.8504	.1260	.9418	.9448	.8510	.8514	.0177	.9472	.9456	.8564	.8575	.0002	.9450
		(.7805)	(.8761)	(.8750)	(.1038)	(.5484)	(.9791)	(.9914)	(.9914)	(.0890)	(.9253)	(.9999)	(.9999)	(.9999)	(.0940)	(.9992)
160	1	.9504	.9521	.9537	.8836	.9515	.9493	.9498	.9502	.8199	.9506	.9492	.9496	.9502	.7272	.9500
		(.9964)	(.9964)	(.9964)	(.9910)	(.9780)	(1.0)	(1.0)	(1.0)	(1.0)	(.9999)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
160	2*d*	.9493	.9782	.9788	.9349	.9514	.9472	.9767	.9774	.9282	.9472	.9462	.9775	.9779	.9269	.9497
		(.9964)	(.9953)	(.9953)	(.9994)	(.9784)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
160	2*i*	.9447	.9013	.9037	.6624	.9453	.9518	.9108	.9114	.4841	.9502	.9536	.9097	.9108	.2569	.9520
		(.9892)	(.9926)	(.9925)	(.8339)	(.9477)	(1.0)	(1.0)	(1.0)	(.9732)	(.9994)	(1.0)	(1.0)	(1.0)	(.9997)	(1.0)
160	4*d*	.9467	.9929	.9929	.8344	.9512	.9489	.9915	.9915	.7386	.9477	.9518	.9914	.9917	.5832	.9557
		(.9830)	(.9622)	(.9622)	(.9996)	(.9392)	(1.0)	(.9998)	(.9997)	(1.0)	(.9995)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
160	4*i*	.9453	.8528	.8570	.3058	.9505	.9489	.8590	.8594	.1093	.9497	.9444	.8507	.8525	.0150	.9447
		(.9451)	(.9754)	(.9754)	(.3235)	(.8310)	(.9995)	(.9999)	(.9999)	(.3215)	(.9968)	(1.0)	(1.0)	(1.0)	(.3938)	(1.0)

Note. The ratio between sample sizes is constant in the k studies of the same meta-analysis. π_T and π_C = population proportions, k = number of studies, $\overline{N}$ = average sample size, n_T/n_C = ratio between sample sizes, and $\delta = \pi_T - \pi_C$ = population risk difference.

a. *d* denotes a direct relationship where the most extreme proportion is associated with the lowest sample size.

b. *i* denotes an inverse relationship where the most extreme proportion is associated with the largest sample size.

inverse relationship, d_{CW} estimator presented the values .2425 and .0177, for $n_T/n_C = 2$ and 4, respectively.

Statistical Power

The last criterion in the comparison of the five procedures was the statistical power of the confidence interval to reject the hypothesis of null effect, H_0: $\delta = 0$. Tables 8 and 9 show (in parentheses) the empirical power of the five procedures in conditions in which we have mixed different ratios in sample sizes throughout each of the meta-analyses. As expected, the power of all of the procedures increased as the number of studies, k, the average sample size, $\overline{N}$, and the population risk difference, δ, increased. As also expected, the more extreme the population proportions, the higher the power.

The confidence intervals over the d_C and d_{MH} estimators showed a systematically larger power with an inverse sample size-population proportion relationship. In contrast, the remaining procedures (d_{ML}, d_{CW}, and d_U) presented the opposite trend: a systematically larger power with a direct relationship rather than with an inverse one. It is interesting that this effect was clearly larger for the d_{CW}'s confidence interval. For example, with $k = 10$, $\overline{N} = 100$, and $\pi_T = 0.125$ and $\pi_C = 0.075$, the empirical power for the d_{CW}'s confidence interval was 0.8632 and 0.4067 in meta-analyses with a direct and an inverse relationship, respectively (see Table 9).

As expected, when the n_T/n_C ratio increased from equal sample sizes to unbalanced ones, the power decreased in the confidence intervals of the d_{ML}, d_C, d_{MH}, and d_U estimators. As an example, Table 10 presents the empirical power (in parentheses) of the five procedures for the most extreme population proportions ($\pi_T = 0.075$ and $\pi_C = 0.025$). However, the d_{CW} procedure showed a different trend as a function of the sample size-population proportion relationship. In particular, with an inverse relationship, the larger the ratio between sample sizes, the lower the power; whereas with a direct relationship, as the ratio between sample sizes increased, the power also increased.

As a general rule, the d_{ML}'s confidence interval presented the largest empirical power values, closely followed by the d_C and d_{MH} procedures, the d_C procedure showing a slightly larger power than that of d_{MH}. The d_U procedure showed the lowest empirical power values, and the d_{CW} procedure exhibited a very irregular performance, even showing the largest and lowest power values as a function of the manipulated conditions.

In some conditions, the d_{ML} procedure did not achieve the largest power, its power being lower than that of the d_{CW} procedure in some cases or lower than that of both d_C and d_{MH} procedures in other cases. However, this better power was at the expense of an unadjusted empirical confidence level in the procedure with a higher power than that of d_{ML}. For example, with $k = 10$, $\overline{N} = 100$, $\pi_T = 0.125$, $\pi_C = 0.075$, and a direct relationship between sample sizes and the

population proportions, the power values for d_{CW}, d_{ML}, d_C, d_{MH}, and d_U were 0.8632, 0.7264, 0.6628, 0.6585, and 0.5384, respectively; thus, although d_{CW} showed the highest power, it was at the cost of an unadjusted empirical confidence level of 0.8812; the remaining procedures, however, conformed to the 0.95 nominal confidence level (see Table 9).

As mentioned previously, the dependence between the sample risk differences and the estimated weights explains the irregular performance of the d_{CW} procedure. The most problematic conditions are those of a very low statistical power, which are obtained when the population proportions are in an extreme position and the most extreme proportion is associated with the largest sample size. In these conditions, sample risk differences with a different sign (plus or minus) to that of the population effect size received a disproportionate weight, leading to the erroneous acceptance of the null hypothesis.

Conclusions

This article compared the performance of five estimators of a common risk difference in a meta-analysis composed of a set of k independent 2×2 contingency tables. Our simulations assumed that the set of studies would estimate a common population risk difference, δ, that is, homogeneity in the risk difference was assumed. Furthermore, we also assumed that all of the treatment groups would estimate a common population proportion, π_T, and the same would be the case for the control groups, all of them estimating the same population proportion, π_C. The comparison criteria for the five estimators were bias, relative efficiency, confidence level adjustment, and statistical power. Thus, the focus of our work is limited to the first of the three above-mentioned purposes in a meta-analysis—to estimate the average effect size.

Under the simulated conditions, the d_{ML}, d_C, d_{MH}, and d_U estimators are unbiased. The exception is the d_{CW} estimator, showing a slight negative bias. Therefore, the key difference between the estimators is their variability or, in other words, their relative efficiency. The difference in variability explains the differential performance of the estimators in confidence interval adjustment and statistical power.

Taking into account all of the comparison criteria in our simulation, d_{ML} is the best unbiased estimator of a common risk difference. In fact, the d_{ML} confidence interval systematically shows a good adjustment to the nominal confidence level. Its statistical power for testing the null hypothesis H_0: $\delta = 0$ is the highest in most of the manipulated conditions. Furthermore, when the other estimators (d_C, d_{MH}, and d_{CW}) present a higher power than that of the d_{ML} estimator, it is at the expense of confidence intervals that do not conform to the nominal confidence level. Nevertheless, the power was clearly insufficient in many of the conditions, being lower than the 0.80 value Cohen (1988) recommended.

On the other hand, the d_C and d_{MH} estimators exhibit a very similar performance to one another through the comparison criteria. Although they do not conform to the nominal confidence level under certain conditions, the d_C and d_{MH} estimators present an efficiency and statistical power close to that of the d_{ML} estimator. Therefore, caution is advised in applying the d_C and d_{MH} estimators under such problematic conditions, that is, when there is a direct or inverse relationship between the sample sizes and the population proportions.

Although unbiased and with a good adjustment to the nominal confidence level, the d_U estimator shows an inferior efficiency and statistical power than that of the d_{ML}, d_C, and d_{MH} estimators. This is because the d_U does not apply any weighting scheme to the individual risk differences, ignoring the information of binominal distributions. Due to the low statistical power of the d_U estimator, its application in practice is not recommended.

Apart from its slightly negative bias, the d_{CW} estimator shows an irregular performance in all of the comparison criteria. Our results lead us to advise against the use of such an estimator, especially in meta-analyses with low sample sizes, a large number of studies, unbalanced sample sizes, extreme population proportions, and an inverse relationship between sample sizes and population proportions. The dependence between risk differences and estimated weights explains the erratic performance of the d_{CW} estimator, especially when there are zero values in the 2×2 tables. The d_{CW} estimator would improve its performance if it were computed by estimating π_T and π_C in Equation 11 from a common estimate for all the studies after adding together all the k 2×2 tables, instead of using the individual proportions of each study, p_{T_i} and p_{C_i}. Nevertheless, we preferred to compute the d_{CW} estimator such as it is usually proposed and applied (e.g., Shadish & Haddock, 1994). On the other hand, the performance of the d_{CW} estimator could also improve by assuming a random-effects model instead of the fixed-effects model we have considered (e.g., see DerSimonian & Laird, 1986; Emerson et al., 1993, 1996). This would be so because under a random-effects model the variance of each risk difference would include a constant, the between-studies variance component. Thus, the greater the between-studies variance, the lower the dependence between the risk differences and their weights. In any case, it would be of interest to study the performance of the proposed modifications.

Assuming complete homogeneity, the results of our simulation study confirm the usefulness of the risk difference to average the effect sizes in a series of 2×2 tables. Under these conditions, we advise choosing the risk difference as the effect size measure, although always reporting the average risk difference together with the average proportions in treatment and control groups. This analytic strategy will offer a very clear picture of the evidence.

Although under complete homogeneity assumption the d_{ML} index is the estimator of choice, it is very common in practice to find meta-analyses

where such a restrictive assumption is not met. When all of the empirical studies in a meta-analysis estimate a common risk difference, but there are variations among the treatment population proportions and/or the control population proportions, d_C and d_{MH} would be more advisable estimators than d_{ML}. This is because, unlike the d_{ML} estimator, the d_C and d_{MH} estimators were designed for situations where the actual proportions may vary from study to study, the risk differences remaining homogeneous. That is to say, d_C and d_{MH} estimators do not require the complete homogeneity assumption, but only the homogeneity of risk difference assumption. Consequently, even with a slightly worse performance than that of the d_{ML} estimator, the d_{MH} and d_C estimators have the advantage of a wider area of application. In any case, simulation studies where the control and treatment rates vary slightly through the studies could still be of interest if we wished to compare the performance of the d_{ML}, d_{MH}, and d_C estimators in less restrictive conditions than those of complete homogeneity.

References

Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology*, *28*, 369-404.

Aptech Systems. (1992). *The Gauss System version 3.0*. Kent, WA: Author.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, *10*, 417-451.

Cochran, W. G. (1983). *Planning and analysis of observational studies*. New York: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

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

Cooper, H. M. (1989). *Integrating research: A guide for literature reviews* (2nd ed.). Beverly Hills, CA: Russell Sage Foundation.

DerSimonian, R., & Laird, N. M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177-188.

Emerson, J. D., Hoaglin, D. C., & Mosteller, F. (1993). A modified random-effect procedure for combining risk difference in a set of 2×2 tables from clinical trials. *Journal of the Italian Statistical Society*, *3*, 269-290.

Emerson, J. D., Hoaglin, D. C., & Mosteller, F. (1996). Simple robust procedures for combining risk differences in sets of 2×2 tables. *Statistics in Medicine*, *15*, 1465-1488.

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: Russell Sage Foundation.

Goldman, L., & Feinstein, A. R. (1979). Anticoagulants and myocardial infarction. *Annals of Internal Medicine*, *90*, 92-94.

Greenland, S. (1987). Quantitative methods in the review of epidemiologic literature. *Epidemiologic Reviews*, *9*, 1-30.

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., & McCrory, D. C. (1995). Meta-analytic tools for medical decision making: A practical guide. *Medical Decision Making*, *15*, 81-96.

Hasselblad, V., Mosteller, F., Littenberg, B., Chalmers, T. C., Hunink, M.G.M., Turner, J. A., Morton, S. C., Diehr, P., Wong, J. B., & Powe, N. R. (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.

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Petitti, D. B. (2000). *Meta-analysis, decision analysis, and cost-effectiveness analysis: Methods for quantitative synthesis in medicine* (2nd ed.). New York: Oxford University Press.

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

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.

Rothman, K. J. (1986). *Modern epidemiology.* Boston: Little, Brown.

Sánchez-Meca, J., & Marín-Martínez, F. (1998). Testing continuous moderators in meta-analysis: A comparison of procedures. *British Journal of Mathematical and Statistical Psychology*, *51*, 311-326.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*, 529-540.

Shadish, W. R. (1996). Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. *Psychological Methods*, *1*, 47-65.

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.

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

Vacha-Haase, R. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6-20.

Whitehead, A., & Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine*, *10*, 1665-1677.

Whitehead, J. T., & Lab, S. P. (1989). A meta-analysis of juvenile correctional treatment. *Journal of Research in Crime and Delinquency*, *26*, 276-295.