British Journal of Mathematical and Statistical Psychology (1998), 51, 311–326 *Printed in Great Britain*

Testing continuous moderators in meta-analysis: A comparison of procedures

Julio Sánchez-Meca† and **Fulgencio Marín-Martínez**

Departamento Psicología Básica y Metodología, Facultad de Psicología, Campus de Espinardo, Apdo 4021, 30080-Murcia, Spain

Using Monte Carlo simulation, we compare the statistical power and Type I error rates of three procedures commonly applied in meta-analysis to detect the influence of a continuous moderator variable on the estimated effect sizes from a set of independent primary studies. We use the conventional T test (TTEST), the Hedges & Olkin (1985) z_{HO} test (H&O), and the Rosenthal & Rubin (1982) z_{RR} test (R&R). The H&O and R&R procedures weight every estimated effect size by its inverse-variance, whereas in the conventional T test the individual effect sizes are unweighted. Results show an adequate control of the Type I error rate for the three procedures in all of the conditions. The statistical power is very similar in the H&O and R&R procedures, and systematically lower in the conventional T test. Both H&O and R&R procedures were similarly the most appropriate and powerful to detect a quantitative moderator variable on effect sizes, supporting the suitability of applying weighting schemes in meta-analysis.

1. Introduction

In the last 20 years, meta-analysis has become a popular methodology for quantitatively integrating the results of several studies about the same research topic. The purpose of meta-analysis can be summarized in three main objectives: (a) to average the results of the studies, (b) to assess the homogeneity of the results, and (c) to search for moderator variables to explain the heterogeneity of the results. An effect size index is usually chosen as the standard measure to express the studies' results, enabling them to be compared with each other (Becker, 1988; Cooper & Hedges, 1994; Glass, McGaw & Smith, 1981; Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Johnson & Eagly, in press; Rosenthal, 1991).

The representativeness of the average effect size of a set of studies depends on the degree of heterogeneity in the pooled effect sizes. Specifically, the average effect size is a conclusive indicator of the estimate of the strength of the relationship only under the assumption of a common population estimate. On the other hand, as the heterogeneity of the effect sizes increases, the interpretation of an average effect size becomes more and more complex, although it can give a useful indication of the general trend in the results.

There are different statistical significance tests to analyse whether the sample effect sizes

† Requests for reprints; email (jsmeca@fcu.um.es).

estimate a common parameter or, alternatively, the differences among effect sizes are so large that they cannot be explained by sampling error. Nevertheless, several Monte Carlo studies have shown that the statistical power of these tests is not high enough, and that acceptance of the effect sizes homogeneity hypothesis may often lead to Type II error (Rassmussen & Loher, 1988; Sackett, Harris & Orr 1986; Sagie & Koslowsky, 1993; Sánchez-Meca & Marín-Martínez, 1997; Spector & Levine, 1987). For example, Sánchez-Meca & Marín-Martínez (1997) found that, in many of the manipulated conditions, the power of several statistical tests was lower than the 0.80 value Cohen (1988) recommended, with the empirical values averaging around 0.35.

To interpret the heterogeneity of effect sizes, meta-analysis formulates statistical models in which it is possible to explain such heterogeneity as a function of the influence of substantive and methodological characteristics of the primary studies (Lipsey, 1994). As an example, variables such as methodological quality of the studies, publication year, selected sample, or several operational definitions for the same construct can moderate the sample effect sizes variability. The general linear model for predicting the effect sizes from moderator variables is the usual strategy for analysing their possible association. Two major meta-analytic approaches widely applied in practice propose different statistical tests to analyse the influence of a quantitative moderator variable (Abrami, Cohen & d'Apollonia, 1988; Bangert-Drowns, 1986; Johnson, Mullen & Salas, 1995; Marín-Martínez, 1996); they are developed in Hedges & Olkin (1985) and in Rosenthal & Rubin (1982; see also Rosenthal, 1991). These tests have been devised specifically to be applied to meta-analytic data (for example, effect sizes). Concretely, these procedures assume differing accuracy in the studies' effect sizes, weighting each one by its inverse-variance; thus, these procedures are based on estimation by weighted least squares. Examples of the use of these procedures are the meta-analyses of Eagly & Johnson (1990) who follow Hedges & Olkin, and of Mullen, Johnson & Salas (1991) who apply Rosenthal & Rubin.

An alternative practice consists of applying a conventional statistical test, computing the product–moment correlation coefficient between the continuous moderator variable and the effect sizes, and testing its statistical significance by means of the conventional T test (cf., for example, the meta-analysis of Shapiro, Harper, Startup, Reynolds, Bird & Suokas, 1994). In this way, all of the effect sizes are equally weighted, obviating their differences in accuracy. Hunter & Schmidt (1990, p. 408) contend that the conventional tests are usually robust with respect to violations of the homoscedasticity assumption, rendering unnecessary the application of special techniques in meta-analysis, such as weighted least squares.[1] On the other hand, some authors expressly advise against the use of conventional statistical methods in meta-analysis, such as T tests, analysis of variance, or regression analysis (Abrami *et al.*, 1988; Durlak & Lipsey 1991; Hedges & Becker, 1986; Hedges & Olkin, 1985; Johnson & Eagly, in press; Johnson & Turco, 1992; Rosenthal, 1991).

The meta-analytic results from these different procedures will not necessarily coincide. (In fact, Johnson *et al.* (1995) applied these three approaches to the same meta-analytic database and observed different results for each procedure.) If alternative procedures to test the influence of a quantitative moderator variable do not give the same results, then meta-analysts applying different procedures to the same database would reach different conclusions.

[1] In any case, Hunter & Schmidt (1990, p. 408) recommend against the use of hypothesis testing in meta-analysis because of the problems of capitalization on chance and low statistical power.

Therefore, examining the statistical properties (for example, Type I error and power rates) of the procedures will help in selecting of the most suitable one in a given meta-analysis.

The procedures proposed from the three approaches depend on large sample approximations to the distributions of effect size estimates and the corresponding statistical tests. Meta-analyses within the behavioural sciences usually involve small sample sizes and a small number of studies; accordingly, each approximation must be evaluated within the particular conditions of typical meta-analyses.

There are no simulation studies on the statistical power and Type I error rates of the three procedures. The only exception is that of Hedges & Olkin (1985, p. 180) who, with respect to the Hedges & Olkin z_{HO} test, reported an adequate control of the Type I error rate, although the conditions of their simulation study were very specific: only six studies, two sample size distributions, and two specific design matrices as moderator variables.

The purpose of this article is to compare three statistical significance tests for analysing the influence of a quantitative moderator variable on effect sizes. These are the conventional T test, the Hedges & Olkin procedure, and the Rosenthal & Rubin procedure. Power and Type I error rates of the three procedures were evaluated using Monte Carlo simulation. The number of studies, sample sizes in the studies, and the distribution of the population effect sizes were some of the manipulated conditions. A set of primary studies with a two-group design (commonly experimental *vs.* control) was assumed.[2]

2. The three procedures

Suppose we have a set of k independent studies, each with a two-group design (experimental *vs.* control) and a continuous outcome variable. For the ith study, the population effect size can be conceptualized as the standardized mean difference δ_i, given by (Hedges & Olkin, 1985, p. 76):

$$\delta_i = \frac{\mu_i^E - \mu_i^C}{\sigma_i}, \tag{1}$$

where μ_i^E and μ_i^C are the means of the experimental and control populations of the ith study, respectively; and σ_i is the common population standard deviation of the ith study.

With $\bar{y}_i^E$ and $\bar{y}_i^C$ as the experimental and control means of the ith study, respectively, an unbiased estimate of δ_i is defined by (Hedges, 1981; Hedges & Olkin, 1985, p. 81):

$$d_i = c(m_i)\frac{\bar{y}_i^E - \bar{y}_i^C}{S_i}, \tag{2}$$

where $c(m_i)$ is a correction factor to get an unbiased estimate of the standardized mean difference, computed as:

$$c(m_i) = 1 - \frac{3}{4(n_i^E + n_i^C) - 9}, \tag{3}$$

[2] The assignment rule of subjects to the two groups can be randomly or non-randomly accomplished. The results of our simulation study are applicable to both experimental (random assignment) or quasiexperimental (non-random assignment) designs (Cook & Campbell, 1979).

and S_i is the estimated within-group standard deviation of the ith study calculated by

$$S_i = \left(\frac{(n_i^E - 1)(s_i^E)^2 + (n_i^C - 1)(s_i^C)^2}{n_i^E + n_i^C - 2} \right)^{\frac{1}{2}}, \tag{4}$$

n_i^E and n_i^C being the sample sizes, and $(s_i^E)^2$ and $(s_i^C)^2$ the unbiased estimates of the variances of the two groups in the ith study.

Let X be a quantitative characteristic, or moderator variable, of the studies with values $x_1, x_2, \ldots, x_j, \ldots, x_J$. Thus, the ith study includes both an X value, X_{ij}, and a standardized mean difference estimate, d_{ij}. The correlational nature of the meta-analytic data implies that the relationship between the X and d variables assumes a bivariate distribution. The purpose of the three procedures shown below is to assess the possible association between the moderator variable, X, and the studies' effect sizes, d. The procedures differ in both the statistical formulae and the selected effect size index.

2.1 Conventional T test

For a set of k independent two-group designs, let the standardized mean difference be the effect size index. Assuming a bivariate normal distribution between the population effect sizes δ_i and the moderator variable X, the conventional T test is based on the Pearson product-moment correlation between the δ and X variables, $\rho_{X\delta}$. The $\rho_{X\delta}$ parameter is estimated by the sample correlation between the pairs d and X values, r_{Xd}, obtained in the k studies. Under H_0: $\rho_{X\delta} = 0$,

$$T = r_{Xd}/s_{r_{Xd}} \tag{5}$$

follows an approximate t distribution with $k - 2$ degrees of freedom (for example, Hays, 1988), $s_{r_{Xd}}$ being the estimated standard error computed via:

$$s_{r_{Xd}} = \left(\frac{1 - r_{Xd}^2}{k - 2} \right)^{\frac{1}{2}}. \tag{6}$$

Thus, absolute values of T that exceed the 100(1-α/2) per cent critical value of the t distribution with $k - 2$ degrees of freedom, will lead to rejection of the null hypothesis of no relationship between the quantitative variable and effect sizes.

2.2 Hedges & Olkin procedure

Hedges & Olkin (1985) proposed the standardized mean difference as the effect size index in meta-analyses where the primary studies are two-group designs. This approach envisages a weighted least squares multiple regression model that, in the case of a single continuous moderator variable, becomes a simple regression model (Hedges, 1982, 1994; Hedges & Olkin, 1983, 1985).

The regression model aims to predict effect sizes (d values) from the moderator variable X, with the inverse-variance of each effect size as the weighting factor. With such an analysis, the influence of the studies' sample size in the model is taken into account, assigning the larger weights to the more accurate or larger samples' estimates.

Let $\mathbf{d} = [d_1, d_2, \ldots, d_k]'$ be a $k \times 1$ vector with k d values; also let $\mathbf{X}$ be the $k \times 2$ design matrix,

with a first dummy vector of ones and the second vector with the X values; in matrix notation the linear model is given by $\mathbf{d} = \mathbf{X}\beta + \epsilon$, where $\beta = [\beta_0, \beta_1]'$ is a 2×1 vector of regression coefficients and $\epsilon = \mathbf{d} - \delta$ is a $k \times 1$ vector of errors. The distribution of ϵ is approximately k-variate normal with mean zero and covariance matrix, $\mathbf{\Sigma}_d$, estimated by $\mathbf{S}_d$, a diagonal covariance matrix:

$$\mathbf{S}_d = \text{diag}[s^2_{d_1}, \ldots, s^2_{d_k}], \tag{7}$$

where the variance of the ith d value is estimated by

$$s^2_{d_i} = \frac{n_i^E + n_i^C}{n_i^E n_i^C} + \frac{d_i^2}{2(n_i^E + n_i^C)}. \tag{8}$$

The vector of regression coefficients, β, is estimated via

$$\mathbf{b} = (\mathbf{X}'\mathbf{S}_d^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}_d^{-1}\mathbf{d}, \tag{9}$$

its covariance matrix being $\mathbf{\Sigma}_\beta$, estimated by

$$\mathbf{S}_b = (\mathbf{X}'\mathbf{S}_d^{-1}\mathbf{X})^{-1}, \tag{10}$$

where $\mathbf{S}_d$ is defined in (7).

To test the null hypothesis of independence between the moderator variable (X) and the effect sizes (d), $H_0 : \beta_1 = 0$, Hedges & Olkin (1985, p. 170; see also Hedges, 1994, p. 296) proposed the z_{HO} test:

$$z_{\text{HO}} = b_1 / S_{b_1}, \tag{11}$$

where b_1 is the regression coefficient estimate of moderator variable, X, obtained from (9), and S_{b_1} is the corresponding standard error obtained from (10).[3]

Under the assumption of a zero regression slope and with large sample sizes, the z_{HO} test follows an approximate standard normal distribution. Thus, absolute values of z_{HO} over the $100(1 - \alpha/2)$ per cent critical value of the standard normal distribution will lead to rejection of the null hypothesis of independence.

2.3 Rosenthal & Rubin procedure

One difference between the Rosenthal & Rubin procedure and the other two procedures is the selected effect size index (Rosenthal, 1991; Rosenthal & Rubin, 1982). Regardless of the design type in the primary studies, Rosenthal & Rubin propose the Pearson correlation coefficient, transformed into Fisher's Z, as the effect size measure. In an experimental study with two groups, the effect size can also be expressed as the point-biserial correlation coefficient r_{pb}, between the dependent variable and the group membership. It is possible to transform the d index (2) into r_{pb} by means of (Hedges & Olkin, 1985, p. 89)

$$r_{pb} = \frac{d}{(d^2 + 4)^{\frac{1}{2}}}. \tag{12}$$

[3] Hedges & Olkin (1985, p. 171; see also Hedges, 1994, p. 297) proposed their more general formula to test the influence of a set of moderator variables: $Q_R = \mathbf{b}'\mathbf{S}_b^{-1}\mathbf{b}$ where $\mathbf{b}$ is the vector of estimated regression coefficients as computed in (9) or a subset of these coefficients, and $\mathbf{S}_b$ is the estimated covariance matrix of the regression coefficients in $\mathbf{b}$ (10). Assuming that $H_0 : \beta_1 = \beta_2 = \ldots = \beta_l$ as true, the Q_R test follows an approximately chi-square distribution with l degrees of freedom.

We obtain Z_r by transforming r_{pb} into Fisher's Z (Rosenthal, 1991, p. 21):

$$Z_r = \frac{1}{2}\log_e\left(\frac{1+r_{pb}}{1-r_{pb}}\right). \tag{13}$$

Thus, Z_{r_i} is an estimate of the population Fisher's Z of the ith study, ζ_i.

Like the Hedges & Olkin procedure, that of Rosenthal & Rubin also considers the sample size of the primary studies, weighting each effect size estimate by its inverse-variance.[4] With Z_{r_i} as the effect size estimate of the ith study, the inverse-variance is given by $w_i = N_i - 3$, where N_i is the sample size of each study, equivalent to the addition of the sample sizes in the two groups ($N_i = n_i^E + n_i^C$).

To test the null hypothesis of no relationship between the moderator variable X and the effect sizes Z_{r_i}, $H_0 : \rho_{X\zeta} = 0$, Rosenthal & Rubin (1982, p. 501; see also Rosenthal, 1991, p. 80) proposed the z_{RR} test:

$$z_{RR} = \frac{\Sigma_{i=1}^{k}\lambda_i Z_{r_i}}{\left(\Sigma_{i=1}^{k}\frac{\lambda_i^2}{w_i}\right)^{\frac{1}{2}}}, \tag{14}$$

where, in order to test a linear relationship, λ_i can be computed as the deviation of the X value in the ith study from the mean of the X values: $\lambda_i = x_i - \bar{x}$. Under the assumption of no influence of X in effect sizes variability and large sample sizes, the z_{RR} test follows an approximate standard normal distribution. Consequently, absolute values of z_{RR} higher than $100(1-\alpha/2)$ per cent critical value of the standard normal distribution will lead to rejection of the null hypothesis of no relationship between the quantitative variable and the effect sizes.

In view of the computational differences among the procedures, some predictions can be made about the trend in their Type I error and power rates. First, from large sample theory, adequate control of the Type I error rates of the three procedures can be expected in every condition. Second, the inclusion of the sample sizes as weighting factor in the Hedges & Olkin and Rosenthal & Rubin procedures would explain the higher power of these procedures compared with that of the conventional T test. Finally, as the number of studies, average sample size, effect sizes variability, and the magnitude of the association between effect sizes and the moderator variable increase, power rates will also increase in the three procedures.

3. Method

The simulation study was programmed in GAUSS (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^2 is the common population variance. Pairs of independent random samples of sizes n^E and n^C were generated from these populations. The parametric effect size, δ, was defined as in (1).

Each pair of generated samples simulated the data in a primary research study, for which the d and Z_r indices (2), (13) were computed. A total of k studies simulating the data of a meta-analysis were generated, which gave k random effect sizes. With J as the number of different X values, a fixed X value was also assigned to each study, so that the procedures to

[4] In the Rosenthal & Rubin approach other weighting schemes, such as study quality, can be applied.

test the influence of a moderator variable could be applied. Thus, a total of J conditional normal distributions of effect sizes were defined, one for each X value, with parameters δ_{x_j} and σ_{x_j}.

To determine the Type I error rate in the three procedures, all of the studies within the same meta-analysis estimated a single population effect size, with a null correlation between X and the effect sizes ($\rho_{X\delta} = 0$). To examine the statistical power, the effect size of the ith study, d_{ij} was generated from a conditional normal distribution with δ_{x_j} as the expected value. The pairs of X_j and δ_j values were selected to be correlated at a fixed magnitude, $\rho_{X\delta}$.

The following parameters were manipulated: (1) the average sample size of each meta-analysis, $\bar{N}$ (being $N = n^E + n^C$ and $n^E = n^C$ for each study), with values 30, 50, 80, and 100; (2) the number of studies, k, with values 10, 20, 40, and 100; (3) the product moment correlation, $\rho_{X\delta}$, between the population effect sizes, δ_{x_j}, and the X moderator variable, with values of 0, 0.30, and 0.60; and (4) with $\rho_{X\delta} \neq 0$, both central tendency and variability of the conditional population effect sizes were independently manipulated generating five different conditional effect sizes distributions.

To generate the sample sizes, N_i, of the k studies in a meta-analysis, some properties of the sample size distribution in 30 real meta-analyses in the behavioural sciences field were assessed. In particular, the Pearson skewness index of the distribution, computed for all the meta-analyses, gave a value of +1.546. In accordance with this value, four vectors of ten Ns each were selected: [12, 18, 22, 22, 24, 24, 28, 30, 48, 72], [32, 38, 42, 42, 44, 44, 48, 50, 68, 92], [62, 68, 72, 72, 74, 74, 78, 80, 98, 122], and [82, 88, 92, 92, 94, 94, 98, 100, 118, 142], all with skewness +1.546, and averaging 30, 50, 80, and 100, respectively. These were the sample size distributions for meta-analyses with 10 studies. To get distributions of 20, 40, and 100 studies, each vector was replicated 2, 4, and 10 times, respectively. The k sample sizes were randomly assigned to the k studies of each meta-analysis.[5]

To simulate meta-analyses with a zero correlation between the moderator X and the population effect sizes ($\rho_{X\delta} = 0$), every study estimated the same population size, $\delta = 0.5$, with the next arbitrary values for the X predictor: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].

With 0.30 or 0.60 as the population correlations between the moderator X and the population effect sizes, the central tendency of the δ_{x_j} was manipulated, with the conditions: low ($\bar{\delta} = 0.25$), medium ($\bar{\delta} = 0.55$), and high ($\bar{\delta} = 0.85$) magnitude. The same was performed for variability, with the conditions: low ($\sigma^2_\delta = 0.023$), medium ($\sigma^2_\delta = 0.092$), and high ($\sigma^2_\delta = 0.367$) variances. The values of X and δ_{x_j} for these conditions are shown in Table 1. These conditions were created on the explicit criterion that all database parameters were held constant, except for the respective parameter being varied. The X vectors were replicated 2, 4, and 10 times, in order to get meta-analyses of 20, 40, and 100 studies, respectively, for all conditions.

For each one of the 4 (sample size) $\times$ 4 (number of studies) $\times$ 11 (effect sizes and X distributions) = 176 conditions defined, 1,000 replications were generated by Monte Carlo

[5] The 30 meta-analyses used to simulate the sample size distributions were selected from psychology journals such as *Clinical Psychology Review, Journal of Applied Psychology, Journal of Consulting and Clinical Psychology, Journal of Educational Psychology, Journal of Personality and Social Psychology, Psychological Bulletin.* The meta-analyses ranged over 1981 to 1994 and reported the sample sizes of a total of 1,160 primary studies. The mean sample size was 79, the median was 56 and the standard deviation 65. The Pearson skewness index was found to be +1.546 and the kurtosis coefficient was +1.605. In order to get greater realism in our simulation study, we selected a series of sample size distributions with similar characteristics to those of the empirical sample size database. The list of meata-analyses is available from the authors.

Table 1. δ and X vectors values in the simulated conditions

$\rho_{X\delta} = 0.30$										$\rho_{X\delta} = 0.60$									
$\sigma^2_\delta = 0.092$ $\bar{\delta} = 0.25$		$\sigma^2_\delta = 0.092$ $\bar{\delta} = 0.55$		$\sigma^2_\delta = 0.092$ $\bar{\delta} = 0.85$		$\sigma^2_\delta = 0.023$ $\bar{\delta} = 0.55$		$\sigma^2_\delta = 0.367$ $\bar{\delta} = 0.55$		$\sigma^2_\delta = 0.092$ $\bar{\delta} = 0.25$		$\sigma^2_\delta = 0.092$ $\bar{\delta} = 0.55$		$\sigma^2_\delta = 0.092$ $\bar{\delta} = 0.85$		$\sigma^2_\delta = 0.023$ $\bar{\delta} = 0.55$		$\sigma^2_\delta = 0.367$ $\bar{\delta} = 0.55$	
X	δ	X	δ	X	δ	X	δ	X	δ	X	δ	X	δ	X	δ	X	δ	X	δ
2	−0.2	2	0.1	2	0.4	2	0.325	2	−0.35	2	−0.2	2	0.1	2	0.4	2	0.325	2	−0.35
6	−0.1	6	0.2	6	0.5	6	0.375	6	−0.15	3	−0.1	3	0.2	3	0.5	3	0.375	3	−0.15
5	0.0	5	0.3	5	0.6	5	0.425	5	0.05	6	0.0	6	0.3	6	0.6	6	0.425	6	0.05
1	0.1	1	0.4	1	0.7	1	0.475	1	0.25	1	0.1	1	0.4	1	0.7	1	0.475	1	0.25
9	0.2	9	0.5	9	0.8	9	0.525	9	0.45	8	0.2	8	0.5	8	0.8	8	0.525	8	0.45
10	0.3	10	0.6	10	0.9	10	0.575	10	0.65	5	0.3	5	0.6	5	0.9	5	0.575	5	0.65
7	0.4	7	0.7	7	1	7	0.625	7	0.85	10	0.4	10	0.7	10	1	10	0.625	10	0.85
3	0.5	3	0.8	3	1.1	3	0.675	3	1.05	7	0.5	7	0.8	7	1.1	7	0.675	7	1.05
4	0.6	4	0.9	4	1.2	4	0.725	4	1.25	4	0.6	4	0.9	4	1.2	4	0.725	4	1.25
8	0.7	8	1	8	1.3	8	0.775	8	1.45	9	0.7	9	1	9	1.3	9	0.775	9	1.45

Note, ρ = population correlation between the continuous variable and the population effect sizes; $\bar{\delta}$ = populaton effect sizes mean; σ^2_δ = population effect sizes variance

simulation. Thus, 176,000 meta-analyses were simulated. Throughout the 1,000 replications of each condition, the three procedures to test the influence of the moderator variable were applied. The criterion for the acceptance *vs.* rejection of the independence hypothesis was adjusted to a nominal two-sided significance level of $\alpha = 0.05$. In conditions of zero correlation between X and δ, the proportion of rejections of the null hypothesis in the 1,000 replications was the actual Type I error rate. In conditions with a non-zero correlation between the moderator variable and the effect sizes, the number of rejections of the null hypothesis was the estimated power.

4. Results and discussion

Table 2 presents Type I error rates for three procedures: conventional Ttest (T TEST), Hedges & Olkin (H&O), and Rosenthal & Rubin (R&R), when applied to meta-analyses where the population effect size was kept constant ($\delta = 0.5$), simulating the independence between the moderator variable, X, and the population effect sizes. Type I error rates are shown as a function of the number of studies (k) and the average sample size ($\bar{N}$) in the meta-analyses. Averaging throughout the conditions, the H&O procedure showed the highest average Type I error rate (0.0469) and was close to R&R (0.0464) and T TEST (0.0453). This order changed over conditions, although the differences among the three procedures were always negligible. As expected from large sample theory, the rates for the three procedures conformed to the nominal significance level, $\alpha = 0.05$, in all of the conditions, with actual values slightly below 0.05.

Tables 3 to 7 show the power values for the three procedures as a function of the magnitude of the correlation ($\rho_{X\delta}$) between the moderator variable and the population effect sizes, the number of studies (k), the average sample size ($\bar{N}$), and changes in the central tendency ($\bar{\delta}$) and variability (σ^2_δ) of the population effect sizes. In many conditions the power of the three procedures was clearly insufficient, being below the minimum value of 0.80 recommended in Cohen (1988). For example, with $k = 20$ and $\bar{N} \leq 50$ it was difficult to find adequate statistical power in any of the three procedures. Over all the conditions, the H&O procedure showed the highest mean power (0.5661), closely followed by R&R (0.5460), although this order changed in the individual conditions. As expected, the T TEST procedure showed the lowest power values in all of the conditions (0.3372). The exception was in conditions

Table 2. Type I Error Rates for the three statistical tests

							$\rho = 0$					
	$k = 10$			$k = 20$			$k = 40$			$k = 100$		
$\bar{N}$	TTest	H&O	R&R	TTest	H&O	R&R	TTest	H&O	R&R	TTest	H&O	R&R
30	0.042	0.047	0.048	0.047	0.048	0.049	0.036	0.044	0.041	0.037	0.037	0.042
50	0.048	0.048	0.045	0.048	0.051	0.046	0.048	0.040	0.046	0.040	0.038	0.036
80	0.048	0.050	0.054	0.041	0.053	0.051	0.051	0.045	0.044	0.055	0.055	0.053
100	0.039	0.046	0.048	0.039	0.039	0.035	0.042	0.035	0.041	0.046	0.047	0.042

Note. ρ = population correlation between the continuous variable and the population effect sizes; k = number of studies; $\bar{N}$ = average sample size; TTest = conventional T test; H&O = Hedges & Olkin procedure; and R&R = Rosenthal & Rubin procedure

Table 3. Power rates for the three statistical tests

		$\bar{\delta} = 0.25$ and $\sigma_\delta^2 = 0.092$					
		$\rho = 0.30$			$\rho = 0.60$		
k	$\bar{N}$	*T*Test	H&O	R&R	*T*Test	H&O	R&R
10	30	0.038	0.084	0.081	0.131	0.286	0.249
	50	0.014	0.138	0.138	0.180	0.451	0.449
	80	0.023	0.225	0.222	0.231	0.677	0.671
	100	0.015	0.254	0.237	0.221	0.764	0.767
20	30	0.063	0.164	0.139	0.302	0.489	0.431
	50	0.075	0.255	0.241	0.468	0.726	0.718
	80	0.066	0.400	0.388	0.608	0.924	0.914
	100	0.038	0.439	0.456	0.665	0.968	0.968
40	30	0.154	0.288	0.259	0.635	0.828	0.747
	50	0.178	0.449	0.452	0.856	0.964	0.957
	80	0.195	0.604	0.600	0.968	0.999	1.0
	100	0.241	0.765	0.769	0.986	0.999	0.999
100	30	0.391	0.608	0.541	0.960	0.991	0.979
	50	0.553	0.846	0.845	1.0	1.0	1.0
	80	0.708	0.956	0.955	1.0	1.0	1.0
	100	0.799	0.992	0.992	1.0	1.0	1.0

Note. $\bar{\delta}$ = population effect sizes mean; σ_δ^2 population effect sizes variance; ρ = population correlation between th continuous variable and the population effect sizes; k = number of studies; $\bar{N}$ = average sample size TTEST = conventional *T* test; H&O = Hedges & Olkin test; and R&R = Rosenthal & Rubin test

reaching the maximum power (100 per cent) for the three procedures, where the ceiling effec obscured the differences among the procedures.

As expected, the power of the three procedures increased as the magnitude of the correlation between the moderator variable, X, and the population effect sizes increased with a slightly more pronounced growth in the T TEST procedure. The power of the H&O and R&R procedures is systematically higher than that of T TEST, although the difference reduces with the larger correlation, $\rho_{X\delta} = 0.60$.

The larger the number of studies, the higher was the power in the three procedures; the differences in power among the procedures was also less. Thus, with a large number o studies and a high correlation between the moderator variable and the population effect sizes the power of T TEST could reach the power of H&O and R&R, with a ceiling effect as tha shown in Tables 3 to 7.

With respect to the influence of the average sample size in the meta-analyses, the powe of the R&R and H&O procedures increased as the sample size increased. The T TEST showed a less pronounced increasing trend than did R&R and H&O, but with exceptions Tables 3 to 7 show that in conditions of low power, that is $\rho_{X\delta} = 0.30$ and $k < 40$, there i no clear trend in the power rates in the T TEST procedure, as a function of the average sampl size.

The central tendency of the population effect sizes did not seem to influence the power o the procedures (see Tables 3–5), while changes in the variability of the effect sizes ha an effect in the expected direction: power rates increased as the variability increased (se

Table 4. Power rates for the three statistical tests

		$\bar{\delta} = 0.55$ and $\sigma^2_\delta = 0.092$					
		$\rho = 0.30$			$\rho = 0.60$		
k	$\bar{N}$	*T*Test	H&O	R&R	*T*Test	H&O	R&R
10	30	0.023	0.100	0.079	0.111	0.269	0.207
	50	0.026	0.163	0.129	0.172	0.419	0.415
	80	0.017	0.216	0.212	0.224	0.666	0.643
	100	0.009	0.276	0.263	0.221	0.733	0.727
20	30	0.071	0.174	0.138	0.290	0.506	0.403
	50	0.058	0.257	0.235	0.459	0.719	0.708
	80	0.067	0.366	0.357	0.594	0.908	0.899
	100	0.059	0.459	0.443	0.672	0.949	0.948
40	30	0.114	0.270	0.222	0.609	0.791	0.718
	50	0.190	0.475	0.449	0.857	0.967	0.952
	80	0.206	0.652	0.631	0.946	0.997	0.994
	100	0.223	0.765	0.737	0.982	0.999	0.999
100	30	0.367	0.585	0.504	0.950	0.992	0.976
	50	0.556	0.839	0.803	1.0	1.0	1.0
	80	0.709	0.965	0.960	1.0	1.0	1.0
	100	0.782	0.989	0.985	1.0	1.0	1.0

Note. $\bar{\delta}$ = population effect sizes mean; σ^2_δ population effect sizes variance; ρ = population correlation between the continuous variable and the population effect sizes; k = number of studies; $\bar{N}$ = average sample size; TTEST = conventional *T* test; H&O = Hedges & Olkin test; and R&R = Rosenthal & Rubin test

Tables 6, 4, and 7). This trend was more pronounced in the H&O and R&R procedures, while the TTEST procedure was less sensitive to these changes.

Summing up, the results of our simulation study showed that the three procedures offer a similar and adequate degree of control of Type I error rates, but differ importantly in their power rates. In this respect, H&O and R&R procedures present a similar trend, with a difference in their power values of no practical importance. However, the TTEST procedure deviates from the other two procedures and reaches the lowest power values in most conditions. In particular, the TTEST procedure shows an approximately average 40 per cent lower power than that of H&O and R&R. For example, with $k = 20$ studies, $\bar{N} = 80$, $\bar{\delta} = 0.55$, $\sigma^2_\delta = 0.092$, and $\rho_{X\delta} = 0.60$, TTEST procedure gives a power of 0.594, whereas H&O and R&R reach 0.908 and 0.899, respectively (see Table 4).

The discrepancy of TTEST with respect to the H&O and R&R procedures increases as the sample size and the population effect size variability increase, and conversely, decreases as both the number of studies and the correlation between the moderator variable and effect sizes increase. For example, with $k = 40$ studies, $\bar{N} = 80$, $\bar{\delta} = 0.85$, $\sigma^2_\delta = 0.092$, and $\rho_{X\delta} = 0.60$, the three procedures produce very similar values of 0.946, 0.994 and 0.996, for the TTEST, H&O, and R&R procedures, respectively (see Table 5).

In order to assess whether our empirical power values conform to what would be expected from large sample theory, the expected power was also derived. To this end, equations (11) (H&O) and (14) (R&R) were applied to the population X, δ, and N values; see Appendix. In all the conditions, the theoretically derived results in both procedures were very close to the

Table 5. Power rates for the three statistical tests

		$\bar{\delta} = 0.85$ and $\sigma_\delta^2 = 0.092$					
		$\rho = 0.30$			$\rho = 0.60$		
k	$\bar{N}$	TTest	H&O	R&R	TTest	H&O	R&R
10	30	0.039	0.105	0.065	0.132	0.272	0.225
	50	0.036	0.132	0.120	0.191	0.463	0.419
	80	0.027	0.201	0.169	0.190	0.613	0.585
	100	0.012	0.254	0.230	0.225	0.716	0.694
20	30	0.066	0.176	0.130	0.317	0.483	0.397
	50	0.083	0.255	0.235	0.455	0.730	0.691
	80	0.050	0.356	0.319	0.612	0.898	0.883
	100	0.049	0.446	0.421	0.690	0.963	0.952
40	30	0.132	0.279	0.228	0.576	0.765	0.668
	50	0.159	0.437	0.376	0.838	0.944	0.931
	80	0.189	0.626	0.588	0.946	0.994	0.996
	100	0.245	0.760	0.719	0.973	0.999	0.998
100	30	0.322	0.555	0.452	0.958	0.992	0.975
	50	0.517	0.819	0.756	0.999	1.0	1.0
	80	0.673	0.958	0.947	1.0	1.0	1.0
	100	0.754	0.990	0.987	1.0	1.0	1.0

Note. $\bar{\delta}$ = population effect sizes mean; σ_δ^2 population effect sizes variance; ρ = population correlation between th continuous variable and the population effect sizes; k = number of studies; $\bar{N}$ = average sample size T TEST = conventional T test; H&O = Hedges & Olkin test; and R&R = Rosenthal & Rubin test

empirical ones. For example, the expected values corresponding to $k = 10, \bar{N} = 30, \rho = 0.3$ in Table 4 were 0.095 and 0.081 for H&O and R&R procedures, respectively, against 0.10 and 0.079 as empirical values. Thus these two procedures performed as expected from larg sample theory, even in meta-analyses with a small number of studies ($k = 10$) and a relativel low average sample size ($\bar{N} = 30$).

With respect to the T TEST procedure, it was not possible to derive an exact theoretica power value for each condition, in the same way as for the other two procedures. T TEST (5 differs from the latter procedures, in that does not include the sample size of the primar studies as a weighting factor. Thus, T TEST ignores the variability of the estimated effect sizes which is mainly a function of the sample size in the primary studies (see (8)). Nevertheless, (5) i affected by the variability of the estimated effect sizes since the variability alters the magnitude o the sample Pearson correlation r_{Xd}. As a consequence, our Monte Carlo simulation is a mor reliable way of estimating the power of this procedure than the theoretical derivation.

On statistical criteria, the H&O and R&R procedures are more adequate for detecting th influence of a quantitative moderator variable in meta-analysis; moreover the difference between the two procedures are negligible. In contrast, T TEST is a less powerful procedur because it does not take into account the effect sizes variability as a weighting factor. Onl with a high correlation with the moderator variable and a large number of studies does the powe of the T TEST procedure approximate to that of H&O and R&R, but this situation is uncommo in practice. Thus the conventional T TEST is not an advisable alternative. In all cases, cautio is recommended in conditions where the power of the three procedures is inadequate.

Table 6. Power rates for the three statistical tests

		$\bar{\delta} = 0.55$ and $\sigma_\delta^2 = 0.023$					
		$\rho = 0.30$			$\rho = 0.60$		
k	$\bar{N}$	*T*Test	H&O	R&R	*T*Test	H&O	R&R
10	30	0.038	0.048	0.051	0.069	0.094	0.082
	50	0.025	0.055	0.054	0.097	0.142	0.130
	80	0.044	0.085	0.076	0.122	0.224	0.224
	100	0.035	0.117	0.109	0.134	0.249	0.244
20	30	0.050	0.067	0.064	0.122	0.151	0.139
	50	0.063	0.070	0.075	0.197	0.276	0.252
	80	0.074	0.138	0.131	0.269	0.407	0.383
	100	0.059	0.143	0.133	0.297	0.468	0.455
40	30	0.079	0.102	0.095	0.220	0.300	0.255
	50	0.096	0.139	0.135	0.372	0.455	0.433
	80	0.147	0.214	0.219	0.566	0.692	0.677
	100	0.157	0.292	0.284	0.625	0.752	0.745
100	30	0.144	0.187	0.154	0.459	0.600	0.511
	50	0.245	0.317	0.309	0.733	0.808	0.776
	80	0.295	0.441	0.421	0.932	0.959	0.957
	100	0.361	0.549	0.532	0.972	0.990	0.986

Note. $\bar{\delta}$ = population effect sizes mean; σ_δ^2 population effect sizes variance; ρ = population correlation between the continuous variable and the population effect sizes; k = number of studies; $\bar{N}$ = average sample size; TTEST = conventional *T* test; H&O = Hedges & Olkin test; and R&R = Rosenthal & Rubin test

Our results are similar to those of Johnson *et al.* (1995), and confirm the advisibility of applying statistical procedures that weight the primary studies as a function of the accuracy of the estimated effect sizes (Abrami *et al.*, 1988; Durlak & Lipsey, 1991; Hedges & Becker, 1986; Johnson & Turco, 1992), A more difficult problem is selecting between the H&O and R&R procedures because of their similar performances. The H&O approach has the advantage of including a test of model specification, that is to say, a test of goodness of fit of the linear model to empirical data. A significant result from this test implies that the model is mis-specified, and that searching for other moderators is advisable. Moreover, the H&O approach enables the influence of a set of moderator variables to be tested through multiple regression analysis, whereas R&R is limited to only one moderator variable. On the other hand, the R&R approach has the advantage of simplicity. Therefore, the selected approach by the meta-analyst must be guided by the objectives of the meta-analysis. In particular, when the meta-analysis aims to formulate an explanatory model with the most relevant moderator variables (Cook *et al.*, 1992), the H&O procedure allows the test of such a complex model. When, on the other hand, the meta-analytic perspective is more exploratory than explanatory and there is a reduced number of moderator variables, the simplicity of the R&R procedure is an advantage.

Our results also help in assessing the reliability of meta-analyses already carried out, after examining the performance of these tests in the particular conditions which we have simulated.

Finally, the ecological validity in our simulation study is limited by the manipulated

Table 7. Power rates for the three statistical tests

		$\bar{\delta} = 0.55$ and $\sigma_\delta^2 = 0.367$					
		$\rho = 0.30$			$\rho = 0.60$		
k	$\bar{N}$	*T*Test	H&O	R&R	*T*Test	H&O	R&R
10	30	0.016	0.327	0.242	0.223	0.751	0.691
	50	0.005	0.446	0.428	0.238	0.947	0.953
	80	0.000	0.609	0.615	0.283	0.998	0.996
	100	0.000	0.731	0.736	0.253	1.0	1.0
20	30	0.059	0.520	0.434	0.670	0.965	0.942
	50	0.022	0.701	0.703	0.843	0.998	1.0
	80	0.012	0.893	0.902	0.942	1.0	1.0
	100	0.015	0.953	0.962	0.964	1.0	1.0
40	30	0.206	0.775	0.717	0.974	0.999	0.998
	50	0.248	0.933	0.941	0.998	1.0	1.0
	80	0.274	0.998	0.998	1.0	1.0	1.0
	100	0.247	0.999	1.0	1.0	1.0	1.0
100	30	0.756	0.990	0.981	1.0	1.0	1.0
	50	0.909	1.0	1.0	1.0	1.0	1.0
	80	0.983	1.0	1.0	1.0	1.0	1.0
	100	0.996	1.0	1.0	1.0	1.0	1.0

Note. $\bar{\delta}$ = population effect sizes mean; σ_δ^2 population effect sizes variance; ρ = population correlation between th continuous variable and the population effect sizes; k = number of studies; $\bar{N}$ = average sample size T Test = conventional *T* test; H&O = Hedges & Olkin test; and R&R = Rosenthal & Rubin test

conditions. In particular, letting $n^E = n^C$ in all of the studies is not a common situation i practice. Furthermore, with real meta-analytic data the three procedures can show greate differences, due to the existence of assumption violations, the possible correlation betwee sample sizes and moderators, widely disparate sample sizes, or the presence of outliers. As consequence, new simulation studies are indicated in order to further assess the performanc of the procedures advocated from different meta-analytic approaches.

Acknowledgments

The authors would like to acknowledge the helpful comments of three reviewers. We would also like t thank Mrs Pilar Martínez-Pelegrín (Departamento de Filología Inglesa) for her revision of the Englis language in the article.

References

Abrami, P. C., Cohen, P. A. & d'Apollonia, S. (1988). Implementation problems in meta-analysis *Review of Educational Research*, **58**, 151–179.

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychologica Bulletin*, **99**, 388–399.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal o Mathematical and Statistical Psychology*, **41**, 257–278.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ Lawrence Erlbaum.

Cooper, H. M. & Hedges, L. V. (Eds) (1994). *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Cook, T. D. & Campbell, D. T. (1979). *Quasiexperimentation: Designs and Analysis Issues for Field Settings*. Chicago, IL: Rand McNally.

Cook, T. D., Cooper, H., Cordray, D. F., Hartman, H., Hedges, L. V., Light, R. J., Louis, T. A. & Mosteller, F. (1992). *Meta-analysis for Explanation: A Casebook*. New York: Russell Sage Foundation.

Durlak, J. A. & Lipsey, M. W. (1991). A practitioner's guide to meta-analysis. *American Journal of Community Psychology*, **19**, 291–332.

Eagly, A. H. & Johnson, B. T. (1990). Gender and leadership style: A meta-analysis. *Psychological Bulletin*, **108**, 233–256.

GAUSS (1992). *The GAUSS System Version* 3.0. Washington: Aptech Systems, Inc.

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

Hays, W. (1988). *Statistics*, 4th ed. New York: Holt, Rinehart & Winston.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, **6**, 107–128.

Hedges, L. V. (1982). Fitting continuous models to effect size data. *Journal of Educational Statistics*, **7**, 245–270.

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds), *The Handbook of Research Synthesis*, pp. 285–299. New York: Russell Sage Foundation.

Hedges, L. V. Becker, B. J. (1986). Statistical methods in the meta-analysis of research on gender differences. In J. S. Hyde & M. C. Linn (Eds), *The Psychology of Gender: Advances through Meta-analysis*, pp. 14–50. Baltimore, MD: Johns Hopkins University Press.

Hedges, L. V. & Olkin, I. (1983). Regression models in research synthesis. *American Statistician*, **37**, 137–140.

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

Hunter, J. E. & Schmidt, F. L. (1990). *Methods of Meta-analysis: Correcting Error and Bias in Research Findings*. Beverly Hills, CA: Russell Sage Foundation.

Johnson, B. T. & Eagly, A. H. (in press). Quantitative synthesis of social psychological research. In H. T. Reiss & C. M. Judd (Eds), *Handbook of Research Methods in Social Psychology*. London: Cambridge University Press.

Johnson, B. T., Mullen, B. & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology*, **80**, 94–106.

Johnson, B. T. & Turco, R. (1992). The value of goodness-of-fit indices in meta-analysis: A comment on Hall and Rosenthal. *Communication Monographs*, **59**, 388–396.

Lipsey, M. W. (1994). Identifying potentially interesting variables and analysis opportunities. In H. Cooper & L. V. Hedges (Eds), *The Handbook of Research Synthesis*, pp. 111–123. New York: Russell Sage Foundation.

Marín-Martínez, F. (1996). *Enfoques meta-analíticos: Un estudio comparativo mediante simulación Monte Carlo* [Meta-analytic approaches: A comparison by Monte Carlo simulation]. Doctoral dissertation, University of Murcia.

Mullen, B., Johnson, C. & Salas, E. (1991). Productivity loss in brainstorming groups: A meta-analytic integration. *Basic and Applied Social Psychology*, **12**, 3–23.

Rasmussen, J. L. & Loher, B. T. (1988). Appropriate critical percentages for the Schmidt and Hunter meta-analysis procedure: Comparative evaluation of type I error rate and power. *Journal of Applied Psychology*, **73**, 683–687.

Rosenthal, R. (1991). *Meta-analytic Procedures for Social Research*, revised ed. Newbury Park, CA: Russell Sage Foundation.

Rosenthal, R. & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, **92**, 500–504.

Sackett, P. R., Harris, M. M. & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A Monte Carlo investigation of statistical power and resistance to Type I error. *Journal of Applied Psychology*, **71**, 302–310.

Sagie, A. & Koslowsky, M. (1993). Detecting moderators with meta-analysis: An evaluation and comparison of techniques. *Personnel Psychology*, **46**, 629–640.

Sánchez-Meca, J. & Marín-Martínez, F. (1997). Homogeneity tests in meta-analysis: A Monte Carlo comparison of statistical power and Type I error. *Quality & Quantity*, **31**, 385–399.

Shapiro, D. A., Harper, H., Startup, M., Reynolds, S., Bird, D. & Suokas, A. (1994). The high-water mark of the drug metaphor: a meta-analytic critique of process-outcome research. In R. L. Russell (Ed.) *Reassessing Psychotherapy Research*, pp. 1–35. New York: Guilford Press.

Spector, P. E. & Levine, E. L. (1987). Meta-analysis for integrating study outcomes: A Monte Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology*, **72**, 3–9.

Received 17 June 1997; revised version received 11 December 1997

Appendix

The (large sample) distribution of the z_{HO} statistic is given by

$$z \sim N\left(\frac{\beta_1}{\sigma_{\beta_1}}, 1\right);$$

then the expected power of z_{HO}, P_{HO}, for each of the defined conditions is given by

$$P_{HO} = 1 - \phi\left(|z_{0.025}| - \frac{\beta_1}{\sigma_{\beta_1}}\right),$$

where $\phi(x)$ is the standard normal cumulative distribution function, $|z_{0.025}|$ is the 100 (0.025) per cent critical value of the standard normal distribution (we assume the two-sided $\alpha = 0.05$), β_1 is the regression slope as computed in (9) from the X and δ variables, and σ_{β_1} is the standard error of β_1 as computed in (10) from the X, δ and N values.

On the other hand, the (large sample) distribution of the z_{RR} statistic is given by

$$z \sim N\left(\frac{\Sigma_{i=1}^{k}\lambda_i\zeta_i}{\{\Sigma_{i=1}^{k}\lambda_i^2/(N_i-3)\}^{\frac{1}{2}}}, 1\right);$$

then the expected power of z_{RR}, P_{RR}, for each of the defined conditions is given by

$$P_{RR} = 1 - \phi\left(|z_{0.025}| - \frac{\Sigma_{i=1}^{k}\lambda_i\zeta_i}{\{\Sigma_{i=1}^{k}\lambda_i^2/(N_i-3)\}^{\frac{1}{2}}}\right),$$

where $\phi(x)$ is the standard normal cumulative distribution function, $|z_{0.025}|$ is the 100 (0.025) per cent critical value of the standard normal distribution, ζ_i is the population δ_i value for the ith study transformed into point-biserial correlation coefficient (12) and then into Fisher's Z (13), and λ_i is defined from X variable as in (14).