Quality & Quantity **31**: 385–399, 1997.

Homogeneity tests in meta-analysis: a Monte Carlo comparison of statistical power and Type I error

JULIO SÁNCHEZ-MECA & FULGENCIO MARÍN-MARTÍNEZ
Dpto Psicología Básica y Metodología, Facultad de Psicología, Campus de Espinardo, Apdo 4021, 30080-Murcia, Spain. E-mail: jsmeca@fcu.um.es.

Abstract. The statistical power and Type I error rate of several homogeneity tests, usually applied in meta-analysis, are compared using Monte Carlo simulation: (1) The chi-square test applied to standardized mean differences, correlation coefficients, and Fisher's *r*-to-*Z* transformations, and (2) *S&H-75* (and 90 percent) procedure applied to standardized mean differences and correlation coefficients. Chi-square tests adjusted correctly Type I error rates to the nominal significance level while the *S&H* procedures showed higher rates; consequently, the *S&H* procedures presented greater statistical power. In all conditions, the statistical power was very low, particularly when the sample had few studies, small sample sizes, and presented short differences between the parametric effect sizes. Finally, the criteria for selecting homogeneity tests are discussed.

Key words: meta-analysis, Monte Carlo, effect size.

1. Introduction

Meta-analysis has become a very popular quantitative review technique in social and behavioural research as numerous meta-analytic studies evidence. It aims at integrating the results of a set of empirical studies on a common topic in order to: (1) obtain a global index of the magnitude of the studied relation, (2) test whether the study results are homogeneous, and (3) identify possible variables or characteristics influencing the results obtained (Cooper, 1989; Cooper and Hedges, 1994; Glass et al., 1981; Hedges and Olkin, 1985; Hunter and Schmidt, 1990; Johnson, 1993; Rosenthal, 1991; Sánchez-Meca and Ato, 1989). Statistical considerations on meta-analytic procedures constitute an important issue. Several techniques have been devised to determine the homogeneity of study findings. Using Monte Carlo simulation, this article focuses on comparing Type I and Type II error rates comparing several homogeneity techniques commonly applied in meta-analysis. If different homogeneity tests present different error rates, then the selection of one of them can affect the conclusions of a meta-analysis.

To summarize the results of the studies, these are measured in effect size

indexes. The commonest indexes of the effect size are classified into two families (see Rosenthal, 1994): the *d* family and the *r* family. The most usual index in the *d* family is the standardized mean difference of Hedges (1981), *g*, defined as the difference between the two group means (usually experimental vs. control) divided by the within-group standard deviation.[1] This index is particularly useful to integrate the results of studies with assignment to groups or treatment levels such as the experimental or quasiexperimental studies. On the other hand, the Pearson product moment correlation, *r*, is the typical effect size index in correlational research. There are formulas to transform *d* into *r* and vice versa (Rosenthal, 1991, 1994) and, thus, meta-analysts can select the suitable effect size estimator.

Meta-analysis usually begins applying a homogeneity test to the effect sizes to test whether the average effect size is representative of all of the studies or, on the contrary, there are discrepancies possibly attributed to moderator variables. Each one of the three most frequently used meta-analytic approaches proposes different procedures to test the hypothesis of homogeneity (Bangert-Drowns, 1986; Johnson et al., 1995): (1) The chi-square test of homogeneity for *d* indexes, proposed by Hedges (1994; Hedges and Olkin, 1985), and particularly applied to the field of experimental research; (2) the chi-square test of homogeneity on Fisher's *Z* transformation of *r* indexes, proposed by Rosenthal (1991) and usually applied to correlational research; and stemming from the Schmidt–Hunter procedures, *S&H*, two basic procedures are proposed (Hunter and Schmidt, 1990): (3) The *S&H-75*, particularly applied to the context of the validity generalization of employment tests, and (4) the chi-square test of homogeneity for *r* indexes, applied to correlational research. Consequently, homogeneity tests differ depending on two criteria: type of procedure (chi-square test vs. 75 percent rule) and effect size index (*d*, *r*, or Fisher's *Z* transformation).

The selection of the effect size family (*d* vs. *r*) is determined by the research design (experimental vs. correlational). Nevertheless, Rosenthal (1991, 1994) advocates for *r* family indexes, even in the context of the experimental research, arguing the generality and simplicity of their interpretation. With respect to *r* family, there is a diversity of opinions about the convenience of transforming the correlation coefficient to Fisher's *Z*. Some authors (e.g., Hedges and Olkin, 1985; Rosenthal, 1991; Shadish and Haddock, 1994) recommend to transform the Pearson correlation coefficient when the parametric correlation is extreme arguing the inadequate adjustment of the *r* distribution to normality. Conversely, Hunter and Schmidt (1990) suggest to apply the untransformed correlation coefficient due to the positive bias of Fisher's *Z* distribution.

The three aforementioned homogeneity chi-square tests present an ap-

proximately chi-square distribution with $k - 1$ degrees of freedom, where k is the number of studies. The general formula is given by Shadish and Haddock (1994: 266) as $Q = \Sigma^k w_i(T_i - \bar{T})^2$, where $T_i = i$th effect size estimate, $\bar{T}$ = weighted average effect size, and $w_i = i$th effect size inverse-variance.

In addition to chi-square tests, another homogeneity test for effect sizes is that devised by Schmidt and Hunter, where the sampling error variance and the observed variance of the effect sizes are compared to each other.[2] With real data, Hunter and Schmidt (1990) propose that the homogeneity hypothesis should be accepted when the sampling error variance is equal or superior to 75 percent of the observed variance. But with computer-simulated data, it is advisable to follow a less conservative criterion, because the influence of other statistical artifacts beyond sampling error are minimized. In agreement with Sackett et al. (1986), we have introduced the 90 percent rule as a criterion for rejecting the homogeneity hypothesis.

Up to the present time, several Monte Carlo simulation studies have been accomplished to know the control that some of these procedures have on Type I and Type II error rates. Most of these studies have used the Pearson product moment correlation, r, as the measure of the effect size focusing particularly on the area of test validity generalization (Alexander et al. 1989; Cornwell, 1993; Cornwell and Ladd, 1993; Osburn et al., 1983; Sackett et al., 1986; Sagie and Koslowsky, 1993; Spector and Levine, 1987). Nevertheless, power and Type I error rate of these homogeneity tests when the effect size index is the standardized mean difference have not been explored yet, with the exception in Hedges (1982).

The purpose of the present study is to perform a Monte Carlo simulation to test statistical power and Type I error rate of seven homogeneity tests: (1) chi-square test applied to standardized mean difference, d (Hedges and Olkin, 1985); (2) chi-square test applied to correlation coefficient, r (Hunter and Schmidt, 1990); (3) chi-square test applied to the Fisher's r-to-Z transformation (Hedges and Olkin, 1985; Rosenthal, 1991); (4) and (5) *S&H-75* and *S&H-90* applied to d indexes; and (6) and (7) *S&H-75* and *S&H-90* applied to r indexes. The current study extends the results of the previously mentioned Monte Carlo papers by combining homogeneity tests (chi-square test and *S&H* procedures) together with effect size indexes (d and r families). In this way, the present authors aim to shed light on the contradictory findings of previous simulations. Furthermore, our research incorporates several homogeneity tests not examined before (i.e., the *S&H* procedure with d index, and the power of chi-square test with d index).

To determine Type I error rate we have generated data so that all the studies estimate the same parametric effect size. For the examination of

statistical power the simplest situation has been assumed: a moderator variable with two levels of different parametric effect size for half of the studies in each simulation. The parameters manipulated are the discrepancy between the two parametric effect sizes (to examine Type II error), the value of the parametric effect size, the number of studies, and the sample size which was kept constant at each simulation.

According to the previous simulation studies, in the present study several findings are expected. First, the *S&H* procedures will show higher power than chi-square tests at the expense of higher Type I error rates. Second, chi-square test with d and Fisher's Z indexes will adequately adjust at nominal significance level. Third, statistical power will increase as number of studies, sample size, and parametric effect size differences increase.

2. Method

The simulation study was programmed in GAUSS (1992). Two normally distributed populations with homogeneous variances were defined, $[N(\mu^E, \sigma^2), N(\mu^C, \sigma^2)]$, where μ^E and μ^C are the experimental and control population means, respectively; and σ^2 is the common population variance. From these, pairs of independent random samples were generated with n^E and n^C as sample sizes. The simulated studies were accomplished with an experimental and a control group. The parametric effect size, δ, was defined as (Hedges and Olkin, 1985: 76):

$$\delta = \frac{\mu^E - \mu^C}{\sigma}. \tag{1}$$

Each pair of generated samples simulated the data in a primary research. For each simulated primary study, the following computations were carried out:

1. Means, $\bar{y}^E$ and $\bar{y}^C$, and unbiased variances, $(s^E)^2$ and $(s^C)^2$, of the two samples.

2. Standardized mean difference of Hedges, g, defined as (Hedges and Olkin, 1985: 78):

$$g = \frac{\bar{y}^E - \bar{y}^C}{s}, \tag{2}$$

where s is calculated by:

$$s = \sqrt{\frac{(n^E - 1)(s^E)^2 + (n^C - 1)(s^C)^2}{n^E + n^C - 2}}. \tag{3}$$

3. The bias of the standardized mean difference, g, was corrected via (Hedges and Olkin, 1985: 81):

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

where $c(m)$ is:

$$c(m) = 1 - \frac{3}{4(n^E + n^C) - 9}. \tag{5}$$

4. The d index was transformed into the point-biserial correlation coefficient, r, via (Hedges and Olkin, 1985: 89):

$$r = \frac{d}{\sqrt{d^2 + \frac{4(n-1)}{n}}}. \tag{6}$$

where $n = n^E = n^C$.

5. Transformation of r into Fisher's Z (Hedges & Olkin 1985: 120).

In this way, three indexes of the effect size were obtained for each simulated study: standardized mean difference (4), correlation coefficient (6), and its transformation into Fisher's Z.

Then, a set of k studies simulating the data of a meta-analysis were generated. The following parameters were manipulated: (1) the sample size of each study, $N = n^E + n^C$ (with $n^E = n^C$), with values 30, 50, 80, 100, and 200; (2) the number of studies, k with values 6, 10, 20, 40, and 100; (3) to study Type I error rate, the value of the parametric effect size was manipulated following Cohen's criterion of small, medium, and high effect sizes, with values $\delta = 0.2$, 0.5, and 0.8; (4) to study statistical power, half of the k studies were generated so that the parametric effect size was δ_1, and the second half shared the parametric effect size, δ_2. The discrepancy among δ_1/δ_2 was manipulated across the following conditions: 0.8/0.7, 0.8/0.6, 0.8/0.5, 0.8/0.4, 0.8/0.3, 0.5/0.4, 0.5/0.3, 0.5/0.2, 0.5/0.1, 0.5/0.0, 0.2/0.1, and 0.2/0.0.

For each one of the 5 (sample size) × 5 (number of studies) × 15 (parametric effect sizes) = 375 conditions defined, 1,000 simulation runs were

generated. Each replication simulated one meta-analysis. Thus, 375,000 meta-analyses were simulated. The seven homogeneity tests on effect sizes presented in this paper were applied to each one of these replications (or meta-analyses):

1. The Q_d test, applied to d_i values, has a chi-square distribution with $k - 1$ degrees of freedom (Hedges and Olkin, 1985: 153):

$$Q_d = \sum^{k} w_i^d (d_i - \bar{d})^2 \sim \chi^2_{k-1}, \tag{7}$$

where d_i is given by (4) to each one of the k studies; $\bar{d}$ is the mean effect size weighted by the reciprocal of the variances:

$$\bar{d} = \frac{\Sigma^k w_i^d d_i}{\Sigma^k w_i^d}, \tag{8}$$

and w_i^d is the inverse-variance for the ith effect size:

$$w_i^d = \left[\frac{n_i^E + n_i^C}{n_i^E n_i^C} + \frac{d_i^2}{2(n_i^E + n_i^C)} \right]^{-1}. \tag{9}$$

2. The Q_r test, applied to r_i values, has a chi-square distribution with $k - 1$ degrees of freedom (Shadish & Haddock, 1994: 269):

$$Q_r = \sum^{k} w_i^r (r_i - \bar{r})^2 \sim \chi^2_{k-1}, \tag{10}$$

where r_i results from applying (6) to each one of k studies; $\bar{r}$ is the mean effect size weighted by the reciprocal of the variances:

$$\bar{r} = \frac{\Sigma^k w_i^r r_i}{\Sigma^k w_i^r}, \tag{11}$$

and w_i^r is the reciprocal of the estimated variance of each effect size:

$$w_i^r = \frac{n_i^E + n_i^C - 1}{(1 - r_i^2)^2}. \tag{12}$$

3. The Q_z test, applied to Z_i values, has a chi-square distribution with $k - 1$ degrees of freedom (Rosenthal, 1991: 74):

$$Q_Z = \sum^{k} w_i^Z (Z_i - \bar{Z})^2 \sim \chi^2_{k-1}, \tag{13}$$

where Z_i results from transforming r_i into Fisher's Z; $\bar{Z}$ is the mean effect size weighted by the reciprocal of the variances:

$$\bar{Z} = \frac{\Sigma^k w_i^Z Z_i}{\Sigma^k w_i^Z}, \tag{14}$$

and w_i^Z is the reciprocal of the estimated variance of each effect size:

$$w_i^Z = n_i^E + n_i^C - 3. \tag{15}$$

4. The *S&H-75* percent rule, applied to d_i values ($SH75_d$), based in the percentage of sampling error variance, given by (Hunter and Schmidt, 1990: 414):

$$P_d = \frac{S_e^2}{S_d^2} 100, \tag{16}$$

where S_d^2 is the total variance of the effect sizes and S_e^2 the variance accounted by the sampling error; S_d^2 is given by (Hunter and Schmidt, 1990: 285):

$$S_d^2 = \frac{\Sigma^k N_i (d_i - \bar{d})^2}{\Sigma^k N_i}, \tag{17}$$

where $N_i = n_i^E + n_i^C$ in each study, and d_i and $\bar{d}$ are given by (4) and (8), respectively. The variance accounted by sampling error, S_e^2, is given by (Hunter and Schmidt, 1990: 286):

$$S_e^2 = \left[\frac{N-1}{N-3}\right]\left[\frac{4}{N}\left(1 + \frac{\bar{d}^2}{8}\right)\right], \tag{18}$$

where N is the mean sample size of the k studies in the meta-analysis and $\bar{d}$ was given in (8).

5. The *S&H-75* percent rule, applied to r_i values ($SH75_r$), based in the percentage of sampling error variance, reported by (Hunter and Schmidt, 1990: 414):

$$P_r = \frac{S_e^2}{S_r^2} 100, \tag{19}$$

where S_r^2 is the total variance of the effect sizes and S_e^2 the variance accounted by the sampling error. S_r^2 is given by (Hunter and Schmidt, 1990: 100):

$$S_r^2 = \frac{\Sigma^k N_i (r_i - \bar{r})^2}{\Sigma^k N_i}, \tag{20}$$

where $N_i = n_i^E + n_i^C$ in each study, and r_i and $\bar{r}$ are given by (6) and (11), respectively. The variance accounted by sampling error, S_e^2, is given by (Hunter and Schmidt, 1990: 107):

$$S_e^2 = \frac{k(1 - \bar{r}^2)^2}{\Sigma^k N_i}. \tag{21}$$

To these five procedures, those of *S&H* with a less conservative criterion, the 90 percent rule, were added following the recommendations of Sackett et al. (1986) and Spector and Levine (1987) for the meta-analyses of computer simulated data. So, these were applied to d ($SH90_d$) and r ($SH90_r$) indexes, according to the equations (16) and (19), respectively.

The criterion for the acceptance vs. rejection of the homogeneity hypothesis was a significance level of $\alpha = 0.05$ for Q_d, Q_r, and Q_Z tests. For *S&H* procedures, the homogeneity hypothesis was rejected when P_d (or P_r) < 75 (*S&H-75*), and P_d (or P_r) < 90 (*S&H-90*).

3. Results

3.1. *Type I error rate*

Table 1 presents Type I error rates for all the tests applied with $\delta = 0.5$, and Table 2 summarizes the average Type I error rates as a function of number of studies, sample size, and parametric effect size.[3] Q_Z test systematically presented the lowest Type I error rates across conditions (mean value: 0.031) followed by the Q_d test (mean value: 0.044); in both cases the actual Type I error rate was lower than the nominal significance level. Q_r showed a slightly higher Type I error rate than the nominal level (mean value: 0.080). A close review to Table 2 showed that Q_r Type I error rates increased as number of studies increased, and decreased as a function of sample size and

Table 1. Type I error for chi-square homogeneity tests and Schmidt and Hunter procedures

		$\delta = 0.50$						
k	N	Q_d	Q_r	Q_Z	$SH90_d$	$SH75_d$	$SH90_r$	$SH75_r$
6	30	0.050	0.085	0.037	0.217	0.142	0.276	0.180
	50	0.065	0.076	0.052	0.237	0.153	0.265	0.177
	80	0.045	0.051	0.041	0.231	0.152	0.243	0.161
	100	0.051	0.063	0.043	0.243	0.152	0.246	0.156
	200	0.043	0.042	0.037	0.229	0.150	0.220	0.147
10	30	0.035	0.091	0.028	0.215	0.108	0.294	0.164
	50	0.038	0.062	0.030	0.232	0.125	0.264	0.147
	80	0.050	0.077	0.042	0.249	0.152	0.264	0.155
	100	0.041	0.050	0.037	0.273	0.149	0.275	0.153
	200	0.044	0.046	0.039	0.277	0.144	0.255	0.137
20	30	0.040	0.130	0.026	0.231	0.103	0.331	0.156
	50	0.040	0.073	0.024	0.245	0.098	0.287	0.105
	80	0.037	0.063	0.030	0.234	0.105	0.250	0.106
	100	0.042	0.055	0.037	0.234	0.099	0.238	0.098
	200	0.049	0.049	0.042	0.245	0.103	0.226	0.093
40	30	0.036	0.166	0.021	0.201	0.045	0.329	0.088
	50	0.040	0.101	0.025	0.220	0.053	0.271	0.076
	80	0.046	0.076	0.027	0.224	0.063	0.247	0.068
	100	0.045	0.073	0.030	0.240	0.065	0.244	0.061
	200	0.042	0.039	0.026	0.202	0.051	0.180	0.044
100	30	0.030	0.238	0.021	0.131	0.012	0.273	0.036
	50	0.031	0.134	0.017	0.149	0.010	0.221	0.012
	80	0.046	0.092	0.029	0.183	0.012	0.190	0.020
	100	0.040	0.057	0.025	0.159	0.014	0.155	0.012
	200	0.051	0.050	0.029	0.191	0.012	0.151	0.009

Note. k = number of studies; N = sample size; Q_d, Q_r and Q_Z = chi-square tests applied to d, r, and Z indexes; $SH90_d$, $SH90_r$, $SH75_d$, and $SH75_r$ = Schmidt–Hunter 90 or 75 percent rules applied to d or r indexes; δ = parametric effect size.

parametric effect size. In fact, with small sample sizes and a high number of studies, Q_r presented inadmissibly higher Type I error rates. With Q_Z test, Type I error rate slowly decreased as the magnitude of the effect size increased, but with Q_d test Type I error rate was constant throughout the various effect size values. Sample size and number of studies did not seem to influence Type I error rates of Q_d and Q_Z tests; only Q_Z Type I error rate slightly decreased as number of studies increased.

S&H-75 and *S&H-90*, applied to both d and r values, did not adequately control Type I error rate with the following average values: $SH75_r = 0.103$; $SH75_d = 0.090$; $SH90_r = 0.248$, and $SH90_d = 0.222$. However, there was an exception with $k = 100$, where $SH75_r$ and $SH75_d$ procedures presented Type I error rates below chi-square tests (see Tables 1 and 2). $SH75_r$ and $SH75_d$

Table 2. Average Type I Error rates by number of studies (k), by sample size (N), and by parametric effect size (δ)

k	Q_d	Q_r	Q_Z	$SH90_d$	$SH75_d$	$SH90_r$	$SH75_r$
6	0.049	0.062	0.041	0.230	0.147	0.249	0.159
10	0.042	0.064	0.035	0.248	0.133	0.267	0.148
20	0.046	0.076	0.034	0.245	0.103	0.270	0.117
40	0.044	0.089	0.028	0.228	0.056	0.256	0.071
100	0.040	0.112	0.022	0.162	0.010	0.198	0.020
N	Q_d	Q_r	Q_Z	$SH90_d$	$SH75_d$	$SH90_r$	$SH75_r$
30	0.039	0.138	0.027	0.204	0.085	0.302	0.136
50	0.045	0.090	0.031	0.218	0.090	0.261	0.107
80	0.045	0.069	0.034	0.227	0.090	0.237	0.095
100	0.046	0.060	0.035	0.235	0.095	0.235	0.094
200	0.046	0.045	0.033	0.228	0.089	0.206	0.083
δ	Q_d	Q_r	Q_Z	$SH90_d$	$SH75_d$	$SH90_r$	$SH75_r$
0.2	0.045	0.104	0.040	0.221	0.087	0.241	0.093
0.5	0.043	0.082	0.032	0.220	0.091	0.248	0.102
0.8	0.044	0.056	0.024	0.227	0.091	0.255	0.113

Note. Q_d, Q_r and Q_Z = chi-square tests applied to d, r, and Z indexes; $SH90_d$, $SH90_r$, $SH75_d$, and $SH75_r$ = Schmidt–Hunter 90 or 75 percent rules applied to d or r indexes.

procedures showed decreased Type I error rate as k increased, while $SH90_r$ and $SH90_d$ did not show a clear trend. Sample size influenced the *S&H* procedure when applied to r indexes ($SH75_r$ and $SH90_r$). As can be seen, Table 2 shows that the larger the sample size the smaller the Type I error rate. Furthermore, the *S&H* procedure applied to r values increased Type I error rates as the magnitude of the effect size increased (see Table 2). However, when the procedure is applied to d values no interesting changes were observed as a function of sample size and the parametric effect size.

3.2. *Statistical power*

Tables 3 and 4 show the power values for two of the manipulated subpopulation differences; δ_1/δ_2: 0.5/0.4 and 0.5/0.0, respectively.[3] Table 5 summarizes the average power values as a function of number of studies, sample size, and subpopulation differences. *S&H* procedures showed greater power than Chi-square tests. Mean power values for the four *S&H* procedures were 0.586, 0.425, 0.613, and 0.446 for $SH90_d$, $SH75_d$, $SH90_r$, and $SH75_r$, respectively. For Q_d, Q_r, and Q_Z procedures, mean power values were 0.360, 0.423, and 0.333, respectively. Only for $k = 100$ chi-square tests presented greater power than *S&H-75* ($SH75_d$ and $SH75_r$). Comparing the power of

Table 3. Power values chi-square homogeneity tests and Schmidt and Hunter procedures

				$\delta_1 = 0.50$	$\delta_2 = 0.40$			
k	N	Q_d	Q_r	Q_Z	$SH90_d$	$SH75_d$	$SH90_r$	$SH75_r$
6	30	0.039	0.086	0.034	0.219	0.142	0.274	0.177
	50	0.041	0.060	0.038	0.246	0.142	0.278	0.162
	80	0.056	0.074	0.044	0.269	0.175	0.273	0.181
	100	0.059	0.069	0.050	0.259	0.170	0.268	0.168
	200	0.090	0.092	0.085	0.321	0.230	0.313	0.217
10	30	0.050	0.119	0.045	0.271	0.152	0.338	0.205
	50	0.048	0.080	0.038	0.256	0.149	0.285	0.166
	80	0.063	0.075	0.054	0.291	0.169	0.306	0.171
	100	0.077	0.097	0.071	0.313	0.176	0.311	0.177
	200	0.071	0.072	0.062	0.346	0.209	0.342	0.198
20	30	0.042	0.146	0.033	0.241	0.103	0.328	0.151
	50	0.042	0.102	0.033	0.287	0.111	0.316	0.142
	80	0.065	0.098	0.051	0.353	0.150	0.361	0.164
	100	0.073	0.091	0.056	0.305	0.136	0.319	0.142
	200	0.101	0.107	0.084	0.388	0.188	0.372	0.181
40	30	0.043	0.203	0.027	0.228	0.056	0.342	0.104
	50	0.057	0.130	0.037	0.254	0.066	0.305	0.083
	80	0.075	0.117	0.053	0.311	0.089	0.336	0.099
	100	0.092	0.119	0.073	0.340	0.108	0.339	0.107
	200	0.125	0.135	0.105	0.424	0.150	0.405	0.138
100	30	0.036	0.281	0.017	0.147	0.090	0.287	0.033
	50	0.052	0.199	0.027	0.212	0.013	0.281	0.023
	80	0.079	0.136	0.049	0.260	0.023	0.290	0.025
	100	0.108	0.164	0.077	0.324	0.036	0.323	0.033
	200	0.219	0.234	0.158	0.490*	0.089	0.451	0.077

Note. k = number of studies; N = sample size; Q_d, Q_r and Q_Z = chi-square tests applied to d, r, and Z indexes; $SH90_d$, $SH90_r$, $SH75_d$, and $SH75_r$ = Schmidt–Hunter 90 or 75 percent rules applied to d or r indexes; δ_1 and δ_2 = parametric effect sizes.

tests applied to d values with those to r values, the latter showed larger power than the former. Thus, $SH75_r$ and $SH90_r$ showed larger power than $SH75_d$ and $SH90_d$, and Q_r larger power than Q_d. Nevertheless, this mainly occurred with low sample size and high discrepancy among the two parametric effect sizes.

As expected, the homogeneity tests increased their statistical power as number of studies, k, sample size, N, and the magnitude of the difference among the two parametric effect sizes, δ_1/δ_2, increased (see Table 5). There were, however, some exceptions; when the parametric difference was small (i.e., $|\delta_1 - \delta_2| = 0.1$ or 0.2), Q_r test did not systematically show power values increasing as N increased. Probably, it was due to the deviation from normality in the distribution of sample rs. Moreover, *S&H-75*, applied to d or

Table 4. Power values for chi-square homogeneity tests and Schmidt and Hunter procedures

		$\delta_1 = 0.50$		$\delta_2 = 0.00$				
k	*N*	Q_d	Q_r	Q_Z	$SH90_d$	$SH75_d$	$SH90_r$	$SH75_r$
6	30	0.163	0.249	0.150	0.440	0.320	0.488	0.379
	50	0.325	0.389	0.312	0.663	0.543	0.679	0.565
	80	0.500	0.540	0.490	0.809	0.727	0.817	0.733
	100	0.618	0.649	0.613	0.872	0.805	0.877	0.809
	200	0.932	0.936	0.930	0.992	0.980	0.992	0.980
10	30	0.257	0.418	0.229	0.599	0.442	0.650	0.478
	50	0.402	0.510	0.384	0.727	0.600	0.755	0.622
	80	0.651	0.708	0.635	0.899	0.812	0.905	0.826
	100	0.773	0.813	0.760	0.949	0.900	0.952	0.905
	200	0.988	0.988	0.988	1.0	0.998	1.0	0.998
20	30	0.304	0.571	0.282	0.683	0.469	0.738	0.527
	50	0.629	0.758	0.617	0.882	0.757	0.901	0.774
	80	0.859	0.909	0.848	0.977	0.925	0.979	0.930
	100	0.954	0.965	0.948	0.993	0.976	0.994	0.977
	200	0.999	0.999	0.999	1.0	1.0	1.0	1.0
40	30	0.478	0.775	0426	0.776	0.509	0.840	0.567
	50	0.819	0.921	0.805	0.956	0.837	0.967	0.868
	80	0.992	0.995	0.989	0.999	0.994	0.999	0.995
	100	0.997	1.0	0.997	1.0	0.998	1.0	0.999
	200	1.0	1.0	1.0	1.0	1.0	1.0	1.0
100	30	0.770	0.982	0.710	0.929	0.567	0.973	0.639
	50	0.991	0.998	0.988	0.998	0.967	0.998	0.976
	80	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	100	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	200	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Note. k = number of studies; N = sample size; Q_d, Q_r and Q_Z = chi-square tests applied to d, r, and Z indexes; $SH90_d$, $SH90_r$, $SH75_d$, and $SH75_r$ = Schmidt–Hunter 90 or 75 percent rules applied to d or r indexes; δ_1 and δ_2 = parametric effect sizes.

r values, and with small parametric difference (i.e., 0.1 or 0.2) did not show a clear trend either between power and sample size or between power and number of studies.

4. Discussion

From our findings we can conclude that, in general, the homogeneity tests applied to meta-analysis present insufficient statistical power, that is, lower than the 0.80 value Cohen (1988) recommended. Furthermore, the real power of the analyzed tests is even inferior to that shown in the tables, since the only statistical artifact introduced in our computations was the random

Table 5. Average power values rates by number of studies (k), by sample size (N), and by subpopulation differences ($|\delta_1 - \delta_2|$)

k	Q_d	Q_r	Q_Z	$SH90_d$	$SH75_d$	$SH90_r$	$SH75_r$
6	0.225	0.250	0.210	0.477	0.382	0.494	0.397
10	0.270	0.310	0.252	0.536	0.414	0.558	0.433
20	0.344	0.404	0.321	0.601	0.443	0.627	0.465
40	0.428	0.508	0.396	0.642	0.450	0.675	0.475
100	0.531	0.642	0.486	0.672	0.435	0.713	0.461
N	Q_d	Q_r	Q_Z	$SH90_d$	$SH75_d$	$SH90_r$	$SH75_r$
30	0.154	0.318	0.123	0.395	0.213	0.491	0.285
50	0.258	0.343	0.229	0.510	0.327	0.547	0.359
80	0.369	0.409	0.342	0.605	0.441	0.615	0.450
100	0.427	0.452	0.401	0.650	0.498	0.653	0.500
200	0.590	0.591	0.569	0.767	0.646	0.760	0.638
$\lvert\delta_1 - \delta_2\rvert$	Q_d	Q_r	Q_Z	$SH90_d$	$SH75_d$	$SH90_r$	$SH75_r$
0.1	0.073	0.122	0.056	0.296	0.127	0.321	0.137
0.2	0.208	0.280	0.178	0.478	0.263	0.509	0.279
0.3	0.410	0.481	0.372	0.667	0.488	0.703	0.519
0.4	0.595	0.663	0.565	0.802	0.674	0.830	0.707
0.5	0.732	0.790	0.711	0.884	0.802	0.902	0.828

Note. Q_d, Q_r and Q_Z = chi-square tests applied to d, r, and Z indexes; $SH90_d$, $SH90_r$, $SH75_d$, and $SH75_r$ = Schmidt–Hunter 90 or 75 percent rules applied to d or r indexes.

sampling error. A meta-analysis with real data is affected by other statistical artifacts reducing power, such as unreliability of measurements or range restrictions. Consequently, caution is recommended to use a homogeneity test's nonsignificant result as a criterion to give up searching for moderator variables because of its low statistical power (Hall and Rosenthal, 1991; Johnson and Turco, 1992; Johnson et al., 1995).

The comparison of the various homogeneity tests leads us to several conclusions. First, in a meta-analysis in unfavorable conditions, that is to say, with a small number of studies and small sample sizes, neither power nor Type I error rate are adequately controlled by the *S&H* procedures. On the other hand, chi-square tests, at least, guarantee the control of Type I error rates. Second, choosing among d, r, or Z indexes affects Type I and Type II error rates. Particularly, with r index the power is slightly greater than that of d and Z indexes, but, consequently, the Type I error rate is also greater. If r values are transformed into Fisher's Z, Type I error rates will be below the nominal significance level across conditions. Since the statistical power is low for the chi-square test applied to both r's and Z's indexes, the transformation into Fisher's Z makes possible at least an adequate adjustment at the signifi-

cance level. From this point of view, the findings of the present study support the convenience of transforming r into Fisher's Z.

Generally, there is not any superior homogeneity test. Nevertheless, choosing the most adequate homogeneity test depends on the characteristics of the meta-analysis, such as number of studies, average sample size of the studies and the effect magnitude that is assumed for the moderator variables. Thus, the homogeneity tests proposed by the approaches of Hedges and Olkin (1985), Rosenthal (1991), and Hunter and Schmidt (1990) should be considered on the light of the particular conditions of a meta-analysis. The tables included in this report, as well as those from other Monte Carlo studies, can be very useful to select the most adequate homogeneity test, and to interpret already carried out meta-analyses results.

Acknowledgements

We gratefully acknowledge Dr Blair T. Johnson for his helpful suggestions. Also, we wish to thank (Mrs) Pilar Martínez-Pelegrín (Dpto de Filología Inglesa) for her translation of this report.

Notes

1. The estimator of the standardized mean difference adopted in the present paper is that proposed by Hedges (1981). Cohen (1988) and Glass et al. (1981) propose alternative estimators differring in estimated standard deviation.
2. The Schmidt–Hunter procedure allows to control other statistical artifacts besides sampling error, such as unreliability of measures, range restriction, dichotomization, etc. In our study, these possibilities are excluded because other procedures (chi-square tests) only take into account sampling error. On the other hand, sampling error is the main statistical artifact (Koslowsky and Sagie, 1994).
3. The tables in this report show the general trend in the results. Moreover, Tables 2 and 5 summarize all the results. The remaining tables, excluded because of space reasons, may be requested from the authors.

References

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

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

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

Cooper, H. M. (1989). *Integrating Research: A Guide for Literature Reviews* (2nd edn). Beverly Hills, CA: Russell Sage Foundation.

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

Cornwell, J. M. (1993). Monte Carlo comparisons of three tests for homogeneity of independent correlations, *Educational and Psychological Measurement* 53: 605–618.

Cornwell, J. M. & Ladd, R. T. (1993). Power and accuracy of the Schmidt and Hunter meta-analytic procedures, *Educational and Psychological Measurement* 53: 877–895.

GAUSS (1992). *The GAUSS System* (Vers. 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.

Hall, J. A. & Rosenthal, R. (1991). Testing for moderator variables in meta-analysis: issues and methods, *Communication Monographs* 58: 437–448.

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 categorial models to effect sizes from a series of experiments, *Journal of Educational Statistics* 7: 119–137.

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

Hedges, L. V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. New York: Academic Press.

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

Johnson, B. T. (1993). *DSTAT 1.10: Software for the Meta-Analytic Review of Research Literatures* [manual]. Hillsdale, NJ: Erlbaum.

Johnson, B. T. & Turco, R. M. (1993). The value of goodness-of-fit indices in meta-analysis: a comment on Hall and Rosenthal, *Communication Monographs* 59: 388–396.

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

Koslowsky, M. & Sagie, A. (1994). Components of artifactual variance in meta-analytic research, *Personnel Psychology* 47: 561–574.

Osburn, H. G., Callender, J. C., Greener, J. M. & Ashworth, S. (1983). Statistical power of tests of the situational specificity hypothesis in validity generalization studies: a cautionary note, *Journal of Applied Psychology* 68: 115–122.

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

Rosenthal, R. (1994). Parametric measures of effect size, pp. 231–244 in H. M. Cooper & L. V. Hedges (eds), *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

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. & Ato, M. (1989). Meta-análisis: una alternativa metodológica a las revisiones tradicionales de la investigación [Meta-analysis: a methodological alternative to traditional research reviews], pp. 617–669 in J. Arnau & H. Carpintero (eds), *Tratado de Psicología General. I: Historia, Teoría y Método*. Madrid: Alhambra.

Shadish, W. R. & Haddock, C. K. (1994). Combining estimates of effect size, pp. 261–281 in H. M. Cooper & L. V. Hedges (eds), *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

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.