NJ: Erlbaum.
Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29,* 350–362.
Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika, 34,* 29–35.
Zaremba, S. K. (1965). Note on the Wilcoxon–Mann–Whitney statistic. *Annals of Mathematical Statistics, 36,* 1058–1060.
Zimmerman, D. W. (1996). A note on homogeneity of variance of scores and ranks. *The Journal of Experimental Education, 64,* 351–362.
Zimmerman, D. W., & Zumbo, B. D. (1993). Rank transformations and the power of the Student *t* test and Welch *t'* test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology, 47,* 523–539.
Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology, 51,* 139–149.

The Journal of Experimental Education, 1998, 67(1), 69–81

Testing for Dichotomous Moderators in Meta-Analysis

FULGENCIO MARÍN-MARTÍNEZ
JULIO SÁNCHEZ-MECA
Universidad de Murcia

ABSTRACT. The authors conducted Monte Carlo simulations that compared Type I error rates and the statistical power of 3 tests in detecting the effects of a dichotomous moderator variable in meta-analysis: (a) the Q_B test proposed by L. V. Hedges and I. Olkin (1985), (b) the z_{RR} test proposed by R. Rosenthal and D. B. Rubin (1982), and (c) the z_{HS} test proposed by J. E. Hunter and F. L. Schmidt (1990). Those procedures differ in selected effect size index and in the weighting scheme for each effect size. Number of primary studies, average sample size, and distribution of the parametric effect sizes were the manipulated conditions. The z_{HS} test showed the highest statistical power, although at the cost of an inadequate inflated Type I error rate in meta-analyses with a low number of studies. The Q_B and z_{RR} tests adequately adjusted the Type I error rates, and their statistical power was similar. Criteria for selecting among the tests are discussed.

RESEARCHERS USE META-ANALYSES to integrate quantitatively the results of a set of studies about the same research topic, averaging such results and seeking moderator variables to explain their heterogeneity (Cooper & Hedges, 1994; Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Johnson & Eagly, in press; Rosenthal, 1991). To integrate such results, one should express the primary studies in a common metric, usually as effect size indices (Holmes, 1984; Snyder & Lawson, 1993).

Although there is a common rationale for carrying out a meta-analysis, several approaches differing in their analytic techniques have been proposed (Abrami, Cohen, & d'Apollonia, 1988; Bangert-Drowns, 1986; Johnson, Mullen, & Salas, 1995; Marín-Martínez, 1996). The most commonly applied are those of Hedges and Olkin (1985), Hunter and Schmidt (1990), and Rosenthal and Rubin (1982; Rosenthal, 1991). If different techniques do not yield the same results, then the

selected approach in a given meta-analysis affects its conclusions. In fact, Johnson et al. (1995) obtained different results for the three meta-analytic approaches when they applied the three approaches to the same database.

In the three approaches, one takes the same steps to integrate quantitatively the results of a meta-analytic database: (a) obtain an average effect size; (b) test the homogeneity of the effect sizes; and (c) if the homogeneity hypothesis is not met, search for moderator variables in the primary studies. In the present article, we focused on those statistical tests to determine the influence of a dichotomous moderator variable on effect sizes. Examples of dichotomous moderator variables were publication status of the study (published vs. unpublished), participants' assignment method (random vs. nonrandom), and sample gender (male vs. female).

The discrepancies among the three approaches with respect to their ability to detect a dichotomous moderator variable stem from the selected effect size index for quantifying the results of each study, the weighting factor of the effect sizes, and the specific inferential statistical tests proposed. The statistical properties of such inferential tests have scarcely been studied.

The inferential procedures proposed in the three approaches depend on large-sample approximations to the corresponding statistical tests. However, meta-analyses performed in educational and behavioral sciences usually include small sample sizes and a low number of studies. When statistical theory cannot determine the mathematical properties of statistical tests, one uses simulation studies to explore their performance.

In several simulation studies, the statistical power and Type I error rates of different homogeneity tests have been compared (Harwell, 1997; Rasmussen & Loher, 1988; Sackett, Harris, & Orr 1986; Sagie & Koslowsky, 1993; Sánchez-Meca & Marín-Martínez, 1997; Spector & Levine, 1987). There are no simulation studies in which the procedures have been compared with respect to their ability to test the influence of a dichotomous moderator variable. One exception is the Q_B test of Hedges and Olkin (1985), whose Type I error rate was assessed by Hedges (1982) and by Alliger (1995); they found an adequate adjustment to the nominal alpha level. Alliger also evaluated the statistical power of the Q_B test, although the manipulated conditions were limited to very small samples: meta-analyses with 4 or 10 studies and sample sizes of 10 or 20 participants.

Our purpose in the present study was to use Monte Carlo simulations to compare the statistical power and Type I error rates of three statistical tests in analyzing the influence of a dichotomous moderator variable on effect sizes. Each statistical test was proposed in one of the three aforementioned meta-analytic approaches. The manipulated conditions were number of studies, sample sizes, and distribution of the parametric effect sizes. We assumed a set of independent primary studies with a two-group design (commonly experimental vs. control).

Effect Size Indices

When a primary study has a two-group design (e.g., experimental vs. control) and a continuous dependent variable, the parametric effect magnitude is represented as the population standardized mean difference δ (Hedges, 1981, p. 108):

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

where μ^E and μ^C are the parametric means of the experimental and control populations and σ is the common parametric standard deviation. Hedges (p. 110) proposed to estimate δ by means of the sample standardized mean difference g:

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

where $\bar{y}^E$ and $\bar{y}^C$ are the experimental and control sample means, respectively, and S is the pooled sample standard deviation (Hedges & Olkin, 1985, p. 79; see also Cohen, 1988, p. 67).

Hedges (1981) reported that the g index is a positively biased estimator of the population standardized mean difference δ. To remove the bias on the g index, Hedges (p. 113) defined the unbiased d index as

$$d = c(m)g, \tag{3}$$

where $c(m)$ is a correction factor for obtaining an unbiased estimate of the population standardized mean difference, computed approximately as

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

where n^E and n^C are the experimental and control sample sizes, respectively.

An estimate of the variance of the d index σ_d^2 is given by

$$s_d^2 = \frac{n^E + n^C}{n^E n^C} + \frac{d^2}{2(n^E + n^C)}. \tag{5}$$

Rosenthal (1991) proposed the Pearson correlation coefficient transformed into Fisher's Z as the effect size measure in any meta-analysis, regardless of the design type in the primary studies (i.e., experimental or correlational). Thus, in an experimental study with two groups, the effect size can also be estimated as the point-biserial correlation coefficient r_{pb} between the dependent variable and group membership (Hedges & Olkin, 1985, p. 89). Then, r_{pb} can be transformed into Fisher's Z (Rosenthal, 1991, p. 21).

The Three Procedures

In this article, we assumed a set of k independent studies, each with a two-group design (experimental vs. control), a continuous dependent variable, and a single effect size estimation. Although the same conceptual relationship was tested in all the studies, the effect magnitude can vary as a function of a moderator variable influencing such a relationship. In that respect, we assumed the simplest situation, that is, a dichotomous moderator variable. Thus, the independent effect size estimates fell into two categories, defined a priori by the moderator (grouping) variable. Suppose that there are two disjoint categories of effects, with m_1 effects in the first category and m_2 effects in the second category: $k = m_1 + m_2$. With all the effect sizes in the ith category estimating the same parameter δ_i, one uses the three procedures shown below to test the statistical significance of the difference between the two parametric effect sizes, δ_1 and δ_2.

The Hedges and Olkin Procedure

Hedges and Olkin (1985) proposed the standardized mean difference as the effect size index when the primary studies are two-group designs. Let d_{ij} denote the jth effect size measure in the ith category. Let d_{1+} and d_{2+} be the mean effect estimates for the first and second categories of effect sizes, respectively, where d_{i+} is given as

$$d_{i+} = \frac{\sum_{j=1}^{m_i} w_{ij} d_{ij}}{\sum_{j=1}^{m_i} w_{ij}}. \tag{6}$$

The weight w_{ij} is the reciprocal of the variance of d_{ij}, $w_{ij} = 1/s^2_{d_{ij}}$, where $s^2_{d_{ij}}$ is estimated by Equation 5.

The estimated variance of each d_{i+}, $s^2_{d_{i+}}$, is given by the reciprocal of the sum of the weights in the ith category:

$$s^2_{d_{i+}} = \frac{1}{\sum_{j=1}^{m_i} w_{ij}}. \tag{7}$$

Also, let d_{++} be the grand weighted mean, given by

$$d_{++} = \frac{\sum_{i=1}^{2} \sum_{j=1}^{m_i} w_{ij} d_{ij}}{\sum_{i=1}^{2} \sum_{j=1}^{m_i} w_{ij}}. \tag{8}$$

To test the null hypothesis of no difference between the two parametric effect sizes, Hedges and Olkin (1985, p. 154; see also Hedges, 1982, p. 125, and Hedges, 1994, p. 289) proposed the Q_B test:

$$Q_B = \sum_{i=1}^{2} \frac{(d_{i+} - d_{++})^2}{s^2_{d_{i+}}}, \tag{9}$$

where d_{i+}, $s^2_{d_{i+}}$, and d_{++} are defined in Equations 6, 7, and 8, respectively.

Under the assumption of no difference between the two parametric effect sizes, the Q_B test follows a chi-square distribution with 1 degree of freedom. Thus, values of Q_B higher than $1 - \alpha$ of the critical value of the chi-square distribution with 1 degree of freedom lead to rejection of the null hypothesis of a nonrelationship between the moderator dichotomous variable and the effect sizes.

The Rosenthal and Rubin Procedure

Rosenthal and Rubin (1982) proposed the Pearson correlation coefficient transformed into Fisher's Z as the standard measure of the effect size, with $Z_{r_{ij}}$ representing the jth effect size measure in the ith category. Also, let w_{ij} be the reciprocal of the variance of $Z_{r_{ij}}$, given by $w_{ij} = N_{ij} - 3$, where N_{ij} is the sample size of the jth study in the ith category; that is, $N_{ij} = n^E_{ij} + n^C_{ij}$. To test the null hypothesis of no difference between the two parametric effect sizes, Rosenthal and Rubin (1982, p. 501; see also Rosenthal, 1991, p. 80) formulated the focused z_{RR} test:

$$z_{RR} = \frac{\sum_{i=1}^{2} \sum_{j=1}^{m_i} \lambda_{ij} Z_{r_{ij}}}{\sqrt{\sum_{i=1}^{2} \sum_{j=1}^{m_i} \frac{\lambda^2_{ij}}{w_{ij}}}}, \tag{10}$$

where, to test the difference between two categories of effect sizes, one can compute the λ_{ij} weights as $\lambda_{1j} = 1/m_1$ and $\lambda_{2j} = -1/m_2$, where m_1 and m_2 are the number of effect sizes in each category.

Under the assumption of no difference between the two parametric effect sizes, the z_{RR} test follows a standard normal distribution. Thus, absolute values of z higher than $1 - \alpha/2$ of the critical value of the standard normal distribution lead to rejection of the null hypothesis of a nonrelationship between the dichotomous moderator variable and the effect sizes.

The Hunter and Schmidt Procedure

Hunter and Schmidt (1990) proposed the standardized mean difference as the effect size index in meta-analyses in which the primary studies are two-group designs. With the effect sizes classified in two categories, d_{ij} denotes the jth effect size in the ith category.

Let $\bar{d}_{1.}$ and $\bar{d}_{2.}$ be the average effect sizes for the first and second categories, respectively, where $\bar{d}_{i.}$ is given as

$$\bar{d}_{i.} = \frac{\sum_{j=1}^{m_i} N_{ij} d_{ij}}{\sum_{j=1}^{m_i} N_{ij}}. \quad (11)$$

The variance of the ith mean, $(\sigma_{\bar{d}_{i.}})^2$, can be estimated by (Hunter & Schmidt, 1990, p. 437)

$$(s_{\bar{d}_{i.}})^2 = \frac{\sum_{j=1}^{m_i} N_{ij}(d_{ij} - \bar{d}_{i.})^2}{m_i \sum_{i=1}^{m_i} N_{ij}}. \quad (12)$$

To test the null hypothesis of no difference between the two parametric effect sizes, Hunter and Schmidt (1990, p. 438) proposed the z_{HS} test:

$$z_{HS} = \frac{\bar{d}_{1.} - \bar{d}_{2.}}{\sqrt{(s_{\bar{d}_{1.}})^2 + (s_{\bar{d}_{2.}})^2}}, \quad (13)$$

where $\bar{d}_{1.}$ and $\bar{d}_{2.}$ are defined in Equation 11 and $(s_{\bar{d}_1})^2$ and $(s_{\bar{d}_2})^2$ are the estimated variances of the means, computed by Equation 12. Under the assumption of no difference between the effect sizes in the two groups, the z_{HS} test follows a standard normal distribution. Therefore, absolute values of z higher than $1 - \alpha/2$ of the critical value of the standard normal distribution lead to rejection of the null hypothesis of a nonrelationship between the dichotomous variable and the effect sizes. (Hunter and Schmidt devised the z_{HS} test to be applied in a bare-bones meta-analysis, in which the sampling error is the only statistical artifact considered. Thus, the z_{HS} test does not take into account the influence of other statistical artifacts such as unreliability of measures, range restriction, or dichotomization.)

Summary

We summarize the computational differences among the procedures by describing their respective chosen effect size index and the particular weighting schemes for the effect sizes. In the Hedges and Olkin (1985) and the Hunter and Schmidt (1990) procedures, the standardized mean difference d is used, whereas in the Rosenthal and Rubin (1982) procedure the Pearson correlation coefficient transformed into Fisher's Z, Z_r, is used. In the Hedges and Olkin procedure, each d_{ij} value is weighted by its inverse variance, which is a function of both sample size and the magnitude of the effect size (see Equation 5). Hunter and Schmidt simply weight each d_{ij} value by its sample size (see Equation 11). In the Rosenthal and Rubin procedure, each $Z_{r_{ij}}$ value is weighted by its inverse variance, which is a simple function of the sample size $N_{ij} - 3$.

After showing the computational differences among the procedures, we can make predictions about the trend in their Type I error and power rates. First, we expected to find an adequate adjustment of Type I error rate in the three procedures. Second, in conditions with a low number of studies and small sample sizes, we expected to be able to detect an irregular performance in the procedures. Third, we expected that as the number of studies, the average sample size, and the difference between the two parametric effect sizes increased, statistical power would also increase in the three procedures.

Method

The simulation study was programmed in GAUSS (Aptech Systems, 1992). Two normally distributed populations with homogeneous variances were defined, $[N(\mu^E, \sigma^2), N(\mu^C, \sigma^2)]$, where μ^E and μ^C were the experimental and control population means, respectively, and σ^2 was the common population variance. From those populations, pairs of independent random samples were generated with n^E and n^C as sample sizes. We performed the simulated studies with an experimental and a control group. The parametric effect size δ was defined as in Equation 1.

Each pair of generated samples simulated the data in a primary research study in which d (Equation 3) and Z_r indices were computed. We ran a set of k independent studies simulating the data of a meta-analysis.

To determine the Type I error rate in the three procedures, we estimated a single parametric effect size δ for all of the studies within the same meta-analysis. To examine the statistical power, we generated half of the k studies so that the parametric effect size was δ_1 for the first half and δ_2 for the second half.

We manipulated the average sample size of each meta-analysis $\bar{N}$ (where $N = n^E + n^C$ and $n^E = n^C$ for each study), with values of 30, 50, 80, and 100, and the number of studies k, with $k = 10$, 20, 40, and 100. To study Type I error rate, we manipulated the value of the parametric effect size, following Cohen's (1988) criterion of small, medium, and large effect sizes, with $\delta = 0.2$, 0.5, and 0.8. To study statistical power, we manipulated the discrepancy among δ_1/δ_2 across the following conditions: .5/.4, .6/.5, .5/.2, .8/.5, .5/0, and 1/.5.

To generate the sample sizes N of k studies in a meta-analysis, we assessed some properties of the sample size distribution in 30 real meta-analyses in the field of educational and behavioral sciences. (The list of meta-analyses used to simulate the sample size distribution is available from the authors.) From those 30 meta-analyses, we analyzed the sample size distribution of 1,160 primary studies. The mean sample size was 79, the median was 56, and the standard deviation was 65. We found a positive Pearson skewness index of +1.546 and a kurtosis coefficient of +1.605. To make our simulation study more realistic, we selected a series of sample size distributions with characteristics similar to those

of the empirical sample size database. In accordance with the Pearson skewness index of +1.546, we selected four vectors of 10 sample sizes each: (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 the skewness = +1.546 and averaging 30, 50, 80, and 100, respectively. Those were the sample size distributions for meta-analyses with 10 studies. To obtain meta-analyses of 20, 40, and 100 studies, we replicated each sample size's vector 2, 4, and 10 times, respectively.

For each of the 4 × 4 × 9 (Sample Size × Number of Studies × Parametric Effect Sizes) = 144 conditions defined, we ran 5,000 replications. Thus, 720,000 meta-analyses were simulated. For the 5,000 replications for each condition, we applied the three procedures to test the influence of the moderator variable—the Q_B test (Equation 9), the z_{RR} test (Equation 10), and the z_{HS} test (Equation 13). The criterion for acceptance versus rejection of the independence hypothesis was adjusted to a nominal two-sided significance level of $\alpha = .05$. In conditions with a constant effect size parameter δ for the k studies in the meta-analysis, the proportion of rejections of the null hypothesis in the 5,000 replications was the empirical Type I error rate. In conditions with a dichotomous moderator variable with half of the studies estimating the parameter δ_1 and the other half estimating δ_2, the number of rejections of the null hypothesis was the empirical power.

Results and Discussion

Table 1 contains the Type I error rates for the three procedures in each condition. The Q_B and z_{RR} tests adequately adjusted the Type I error to the nominal significance level. That result corroborates those of Hedges (1982) and Alliger (1995). In contrast, the z_{HS} test showed Type I error rates higher than $\alpha = .05$. Only with a high number of studies ($k \geq 40$) did the z_{HS} test show empirical alpha values close to the nominal alpha level. For example, when $k = 10$, $\bar{N} = 50$, and $\delta = 0.2$, the z_{HS} test had an empirical Type I error rate of .118, significantly larger than the nominal $\alpha = .05$, whereas with a larger number of studies, $k = 100$, $\bar{N} = 100$, and $\delta = 0.8$, the empirical alpha was .050; this adjusted at the nominal $\alpha = .05$ (see Table 1). Neither parametric effect size nor average sample size seemed to influence Type I error rates in any of the three tests. The Q_B and z_{RR} tests were not affected by the number of studies, but in the z_{HS} test, empirical alpha values were adjusted to the nominal significance level as the number of studies increased.

Tables 2 and 3 contain the power values of the three tests. The z_{HS} test obtained the highest power in most conditions, followed by the Q_B and z_{RR} tests, in that order, although the differences between the Q_B and z_{RR} tests were negligible. For example, when $k = 10$, $\bar{N} = 50$, $\delta_1 = 0.5$, and $\delta_2 = 0$, the z_{HS} test achieved a power

TABLE 1
Type I Error Rates for the Three Statistical Tests

$\bar{N}$	$\delta = 0.20$			$\delta = 0.50$			$\delta = 0.80$		
	Q_B	z_{RR}	z_{HS}	Q_B	z_{RR}	z_{HS}	Q_B	z_{RR}	z_{HS}
				k = 10					
30	.042	.047	.117	.042	.044	.119	.043	.042	.119
50	.045	.049	.118	.048	.054	.122	.048	.040	.124
80	.051	.049	.119	.039	.039	.117	.047	.040	.115
100	.047	.050	.117	.047	.047	.117	.047	.041	.122
				k = 20					
30	.045	.051	.078	.044	.046	.073	.044	.043	.078
50	.046	.048	.081	.048	.046	.082	.045	.044	.082
80	.044	.046	.071	.044	.045	.078	.047	.045	.073
100	.046	.048	.074	.049	.048	.082	.045	.039	.077
				k = 40					
30	.043	.048	.063	.043	.046	.061	.044	.040	.062
50	.046	.047	.066	.046	.045	.065	.048	.045	.067
80	.044	.049	.062	.045	.044	.057	.050	.048	.067
100	.044	.046	.061	.048	.049	.065	.047	.042	.065
				k = 100					
30	.043	.048	.054	.046	.052	.058	.044	.042	.057
50	.047	.050	.057	.050	.049	.057	.042	.041	.051
80	.042	.043	.052	.045	.046	.056	.051	.045	.057
100	.044	.044	.055	.047	.045	.054	.043	.037	.050

Note. δ = parametric effect size; k = number of studies; $\bar{N}$ = average sample size; Q_B = Hedges and Olkin test; z_{RR} = Rosenthal and Rubin test; and z_{HS} = Hunter and Schmidt test.

value of .830, whereas the Q_B and z_{RR} tests obtained power values of .772 and .745, respectively (see Table 3).

As expected, the power of the three procedures increased as the number of studies, sample size, and the difference between the two parametric effect sizes increased. Thus, under extreme conditions a ceiling effect operated, achieving similar power values in the three procedures. For example, with $k = 40$, $\bar{N} = 30$, $\delta_1 = 0.5$, and $\delta_2 = 0$, the power values for the Q_B, z_{RR}, and z_{HS} tests were .988, .961, and .990, respectively (see Table 3).

The three procedures presented insufficient statistical power in several of the manipulated conditions, with values lower than the .80 value Cohen (1988) recommended. Under those conditions, one must be cautious in interpreting a nonsignificant result as evidence of nonrelationship. Only with a high number of studies, large sample sizes, or a strong discrepancy between the two parametric

TABLE 2
Power Values for the Three Statistical Tests

	$\delta_1 = 0.50, \delta_2 = 0.40$			$\delta_1 = 0.60, \delta_2 = 0.50$			$\delta_1 = 0.50, \delta_2 = 0.20$		
$\bar{N}$	Q_B	z_{RR}	z_{HS}	Q_B	z_{RR}	z_{HS}	Q_B	z_{RR}	z_{HS}
				k = 10					
30	.058	.061	.123	.057	.054	.121	.238	.202	.353
50	.074	.071	.148	.071	.065	.140	.367	.350	.495
80	.098	.095	.185	.102	.095	.185	.547	.536	.643
100	.123	.118	.204	.114	.110	.201	.630	.621	.713
				k = 20					
30	.083	.078	.121	.076	.065	.117	.421	.356	.496
50	.109	.106	.154	.108	.104	.156	.629	.602	.680
80	.158	.156	.206	.158	.147	.205	.832	.820	.852
100	.191	.189	.247	.180	.175	.228	.915	.907	.926
				k = 40					
30	.119	.108	.153	.119	.096	.147	.697	.596	.732
50	.183	.171	.212	.189	.171	.222	.901	.874	.907
80	.271	.264	.300	.282	.270	.310	.989	.987	.990
100	.330	.318	.364	.332	.321	.367	.998	.998	.998
				k = 100					
30	.255	.213	.281	.242	.193	.267	.978	.942	.981
50	.402	.373	.424	.397	.360	.424	.999	.999	1.0
80	.576	.564	.594	.573	.552	.585	1.0	1.0	1.0
100	.670	.658	.684	.670	.651	.687	1.0	1.0	1.0

Note. δ = parametric effect size; k = number of studies; $\bar{N}$ = average sample size; Q_B = Hedges and Olkin test; z_{RR} = Rosenthal and Rubin test; and z_{HS} = Hunter and Schmidt test.

effect sizes did the procedures show adequate statistical power. Specifically, when the absolute difference between δ_1 and δ_2 was only .10 units, no conditions showed a power of .80 (see Table 2). With an absolute difference between δ_1 and δ_2 of about .30 units, one needs a minimum of 20 studies ($k \geq 20$) to achieve the .80 power value (see Tables 2 and 3). On the other hand, the strongest of the manipulated absolute differences, .50 units, resulted in a high statistical power value in most conditions (see Table 3).

In selecting one of the three tests in a real meta-analysis, researchers must take into account adequate balance between the Type I error and power values, when maximum power and adequate control of the Type I error rates would be desirable. The z_{HS} test was the most powerful of the tests, but at the cost of inflated alpha error in conditions with a low number of studies ($k < 40$). The Q_B and z_{RR} tests, with a similar performance, showed power values lower than the z_{HS} did, but their Type I error rates were always adequately adjusted.

TABLE 3
Power Values for the Three Statistical Tests

	$\delta_1 = 0.80, \delta_2 = 0.50$			$\delta_1 = 0.50, \delta_2 = 0.00$			$\delta_1 = 1, \delta_2 = 0.50$		
$\bar{N}$	Q_B	z_{RR}	z_{HS}	Q_B	z_{RR}	z_{HS}	Q_B	z_{RR}	z_{HS}
				k = 10					
30	.215	.177	.330	.546	.460	.658	.510	.411	.620
50	.360	.332	.474	.772	.745	.830	.765	.712	.824
80	.533	.500	.627	.935	.930	.949	.922	.909	.946
100	.636	.611	.716	.973	.970	.977	.961	.954	.970
				k = 20					
30	.398	.327	.475	.852	.762	.881	.815	.713	.848
50	.626	.573	.674	.970	.960	.975	.962	.942	.967
80	.823	.796	.846	.999	.998	.999	.999	.997	.999
100	.900	.885	.913	1.0	1.0	1.0	1.0	1.0	1.0
				k = 40					
30	.687	.578	.722	.988	.961	.990	.981	.940	.982
50	.893	.865	.906	1.0	1.0	1.0	.999	.999	.999
80	.983	.980	.987	1.0	1.0	1.0	1.0	1.0	1.0
100	.995	.994	.996	1.0	1.0	1.0	1.0	1.0	1.0
				k = 100					
30	.976	.932	.978	1.0	1.0	1.0	1.0	1.0	1.0
50	.998	.998	.998	1.0	1.0	1.0	1.0	1.0	1.0
80	1.0	1.0	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	1.0	1.0

Note. δ = parametric effect size; k = number of studies; $\bar{N}$ = average sample size; Q_B = Hedges and Olkin test; z_{RR} = Rosenthal and Rubin test; and z_{HS} = Hunter and Schmidt test.

In meta-analyses with a low number of studies, the denominator in the z_{HS} test (Equation 13) underestimates the correct sampling variance. As a consequence, the z_{HS} test overestimates its true value, showing both a misleadingly large power and Type I error rates significantly larger than the nominal level $\alpha = .05$. Thus, we advise caution in the application of the z_{HS} test, particularly in meta-analyses with fewer than 40 studies. On the other hand, the application of the z_{HS} test must be considered in the context of the more general Hunter and Schmidt (1990) approach, in which caution is recommended in the use of any hypothesis testing in meta-analysis, because of the problems of capitalization on chance and low statistical power (Hunter & Schmidt, p. 408). Furthermore, the Hunter and Schmidt approach allows one to control other statistical artifacts besides sampling error, such as unreliability of measures, range restriction, and dichotomiza-

tion. In our simulation study, we excluded those possibilities, focusing the comparison of procedures on the sampling error as the only statistical artifact. Although such other artifacts are also important, in practice it is difficult for authors of primary studies to report the information neccessary to control them (Hunter & Schmidt).

The Q_B and z_{RR} tests showed a slightly lower power than the z_{HS} test, although the former two have the advantage of adequately controlling the Type I error rates in all conditions. Therefore, the Q_B and z_{RR} tests are the preferred procedures in general and in meta-analyses with numbers of studies below 40, where the z_{HS} test would not guarantee suitable adjustment of the empirical alpha values. Furthermore, the computational differences between the Q_B and z_{RR} tests, mainly based on the different effect size indices, do not seem to reflect differences in their performance. Thus, a more difficult issue is to select between the Q_B and z_{RR} tests. In that respect, the Hedges and Olkin approach has the advantage of including a test of model specification, that is, a test of the remaining unexplained variability after controlling the influence of the dichotomous moderator variable (Hedges, 1994; Hedges & Olkin, 1985). A significant result in that test implies that the model is misspecified and a search for other moderators is advisable.

Finally, the differences observed in our simulation study are limited to the manipulated conditions. With real meta-analytic data, the three tests could show greater differences because of the existence of assumption violations, the possible correlation between sample sizes and moderators, the widely disparate sample sizes, or the presence of outliers. As a consequence, new simulation studies are needed to further assess the performance of the procedures advocated from different meta-analytic approaches.

NOTE

We would like to thank two anonymous reviewers for their helpful suggestions. Also, we would like to thank Pilar Martínez-Pelegrín (Dpto de Filología Inglesa) for her revision of the English in this article.

Address correspondence to Fulgencio Marín-Martínez, Dpto Psicología Básica y Metodología, Facultad de Psicología, Apdo 4021, Universidad de Murcia, 30100-Murcia, Spain. E-mail: fulmarin@fcu.um.es.

REFERENCES

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

Alliger, G. M. (1995). The small sample performance of four tests of the difference between pairs of meta-analytically derived effect sizes. *Journal of Management, 21,* 789–799.

Aptech Systems. (1992). The GAUSS system (Version 3.0). Maple Valley, WA: Author.

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 ed.). Hillsdale, NJ: Erlbaum.

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

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

Harwell, M. (1997). An empirical study of Hedges's homogeneity test. *Psychological Methods, 2,* 219–231.

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 categorical models to effect size from a series of experiments. *Journal of Educational Statistics, 7,* 119–137.

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285–299). New York: Sage.

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

Holmes, C. T. (1984). Effect size estimation in meta-analysis. *The Journal of Experimental Education, 52,* 106–109.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Beverly Hills, CA: Sage.

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.

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

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* (rev. ed.). Newbury Park, CA: Sage.

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.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *The Journal of Experimental Education, 61,* 334–349.

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.