Averaging Dependent Effect Sizes in Meta-Analysis: a Cautionary Note about Procedures

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

When a primary study includes several indicators of the same construct, the usual strategy to meta-analytically integrate the multiple effect sizes is to average them within the study. In this paper, the numerical and conceptual differences among three procedures for averaging dependent effect sizes are shown. The procedures are the simple arithmetic mean, the Hedges and Olkin (1985) procedure, and the Rosenthal and Rubin (1986) procedure. Whereas the simple arithmetic mean ignores the dependence among effect sizes, both the procedures by Hedges and Olkin and Rosenthal and Rubin take into account the correlational structure of the effect sizes, although in a different way. Rosenthal and Rubin's procedure provides the effect size for a single composite variable made up of the multiple effect sizes, whereas Hedges and Olkin's procedure presents an effect size estimate of the standard variable. The three procedures were applied to 54 conditions, where the magnitude and homogeneity of both effect sizes and correlation matrix among effect sizes were manipulated. Rosenthal and Rubin's procedure showed the highest estimates, followed by the simple mean, and the Hedges and Olkin procedure, this last having the lowest estimates. These differences are not trivial in a meta-analysis, where the aims must guide the selection of one of the procedures.
Key words: meta-analysis, effect size, standardized mean difference, dependence

La estrategia usual para integrar meta-analíticamente los múltiples tamaños del efecto cuando un estudio primario incluye varios indicadores del mismo constructo, es la de promediarlos. En este trabajo se muestran las diferencias numéricas y conceptuales entre tres procedimientos para promediar tamaños del efecto dependientes. Los procedimientos son el de Hedges y Olkin (1985), el de Rosenthal y Rubin (1986) y el de la media aritmética. Mientras que el de la media aritmética ignora la dependencia entre los tamaños del efecto, tanto el de Hedges y Olkin como el de Rosenthal y Rubin tienen en cuenta, aunque de diferente forma, la estructura correlacional de los tamaños del efecto. El procedimiento de Rosenthal y Rubin proporciona el tamaño del efecto de una sola variable compuesta, obtenida a partir de los diversos tamaños del efecto, mientras que el de Hedges y Olkin aporta una estimación del efecto para la variable estándar. Los tres procedimientos se aplicaron a 54 condiciones, manipulándose la magnitud y homogeneidad del vector de los tamaños del efecto y de la matriz de correlaciones entre ellos. Con el procedimiento de Rosenthal y Rubin se obtuvieron las estimaciones más elevadas, seguido del de la media y del de Hedges y Olkin. Estas diferencias no son triviales en un meta-análisis, cuyos objetivos son los que deben guiar la elección de uno u otro de los procedimientos.
Palabras clave: meta-análisis, tamaño del efecto, diferencia media tipificada, dependencia

Meta-analysis has recently become a widely used methodology in the social and behavioral sciences. Its aim is to quantitatively integrate the results of a set of studies about a research problem (Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Marín-Martínez & Sánchez-Meca, 1998; Rosenthal, 1991; Sánchez-Meca & Marín-Martínez, 1997, 1998a, 1998b). The stages to carry out a meta-analysis are similar to those of an empirical study (Cooper, 1989; Rosenthal, 1991; Sánchez-Meca & Ato, 1989): (a) problem formulation, (b) literature review, (c) coding of variables, (d) analysis and interpretation, and (e) publication. The main difference between a primary study and meta-analysis is the unit of analysis; whereas in the former it is (usually) the subject, in the latter it is the study itself. In order to accomplish a meta-analysis, the results of each study have to be expressed in a common metric, capable of representing the relationship between two variables. With this aim, an index of the effect magnitude is defined to represent the overall result of each study. For experimental and correlational studies, the most usual indexes of the effect size (ES) in meta-analysis are the standardized mean difference and the Pearson correlation coefficient, respectively (Rosenthal, 1994).

To guarantee the independence assumption among ESs in a meta-analysis, each empirical study must contribute with just one ES estimate. However, from a single study it is possible to obtain several ES estimates. For example, in a study about gender differences in aggressive behavior, two groups of subjects (male vs. female) are compared on measures such as administering shocks, hitting, and delivering noxious noise; then an ES estimate can be calculated for each measure. If all of the obtained ES estimates are included in the meta-analysis, the independence assumption is not met. Various strategies have been devised to solve this problem: (a) obviate the problem (Glass et al., 1981); (b) select the conceptually most relevant measure (Matt, 1989); (c) carry out separate meta-analyses, one for each measure (Cooper, 1989); (d) include all of the ESs in the meta-analysis, applying appropriate multivariate techniques, such as generalized least square models (Gleser & Olkin, 1994; Kalaian & Raudenbush, 1996; Raudenbush, Becker, & Kalaian, 1988); and (e) average the ESs in each single study contributing with just one estimate in the meta-analysis (Hedges & Olkin, 1985; Marascuillo, Busk, & Serlin, 1988; Rosenthal & Rubin, 1986).

When the outcome measures of a single study are different indicators of the same construct, the usual strategy for overcoming the dependence problem is to average the ES estimates obtained from those measures. Although it is very usual for meta-analysts to calculate a simple mean or median ES to summarize a set of non-independent ESs, these procedures do not take into account the intercorrelations among the outcome measures. However, Marascuillo et al. (1988) uphold calculating a simple average when the set of individual ESs are similar. Two more accurate procedures have been proposed by Rosenthal and Rubin (1986) and Hedges and Olkin (1985; see also Gleser & Olkin, 1994). But the meta-analytic literature has overlooked the conceptual differences between both procedures. In fact, there is not any work dealing with these differences and the existing works referring to both procedures present them as similar (e.g., Abrami, Cohen, & d'Apollonia, 1988; Johnson & Eagly, in press; Kalaian & Raudenbush, 1996; Matt & Cook, 1994; Rosenthal, 1995). After reading these studies, a meta-analyst would erroneously interpret the procedures as interchangeable. The consequences of this confusion are not trivial, because the selection of the procedure for averaging dependent ESs can affect the meta-analysis results. The purpose of this work was to clarify the conceptual and quantitative differences among Rosenthal and Rubin's procedure (*RR*), Hedges and Olkin's procedure (*HO*), and the simple mean procedure. Firstly, the procedures for averaging dependent ESs are presented; then, the estimates obtained by each procedure in several conditions are shown, and, finally, criteria for choosing among the procedures are discussed.

Procedures for averaging dependent ESs

This work focuses on the standardized mean difference, *d*, as the effect size index. It is defined as the difference between two group means (usually experimental vs. control) divided by the within-group standard deviation:[1]

$$d = \frac{\bar{y}^E - \bar{y}^C}{S}, \qquad (1)$$

where $\bar{y}^E$ and $\bar{y}^C$ are the experimental and control group means, respectively, and S is the within-group standard deviation, calculated by:

---

[1] The *d* index proposed in equation (1) is a positively biased estimator of population standardized mean difference, δ (Hedges & Olkin, 1985). The bias can be corrected by the factor *c(m)*:

$$c(m) = 1 - \frac{3}{4m - 9}, \qquad (10)$$

where $m = n^E + n^C$. Thus, an unbiased estimator of *d* is d^u index, given by:

$$d^u = c(m)d. \qquad (11)$$

$$S = \sqrt{\frac{(n^E - 1)(s^E)^2 + (n^C - 1)(s^C)^2}{n^E + n^C - 2}}, \qquad (2)$$

n^E and n^C being the sample sizes, and $(s^E)^2$ and $(s^C)^2$ being the unbiased variances of the two groups.

In a study with p dependent variables measuring the same construct, a vector, $\mathbf{d}$, of d values with p x 1 dimension can be computed, where each d estimate stems from one dependent variable:

$$\mathbf{d'} = (d_1, d_2, ..., d_p) \qquad (3)$$

The associated p x p covariance matrix, $\mathbf{S}$, is given by:

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{12} & ... & s_{1p} \\ s_{21} & s_2^2 & ... & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & ... & s_p^2 \end{bmatrix} \qquad (4)$$

where s_i^2 is the variance of the ith d value and is obtained by:

$$s_i^2 = \frac{n^E + n^C}{n^E \, n^C} + \frac{d_i^2}{2(n^E + n^C)}, \qquad (5)$$

and $s_{ii'}$ $(i \neq i')$ is the covariance between two d estimates, given by:

$$s_{ii'} = r \, s_i \, s_{i'}, \qquad (6)$$

r being the Pearson correlation coefficient between each pair of dependent variables.

The simplest procedure to combine p dependent ESs is an arithmetic mean, ignoring the correlational structure of the data. Thus, we compute the average d value, d_M, according to:

$$d_M = \frac{\sum d_i}{p}, \qquad (7)$$

The inappropriateness of this overall index is obvious because of the dependence of the ESs, but Marascuillo et al. (1988) defend its application if the dependent ESs are homogeneous: "If one is able to pool within-group effect sizes,

it is because the estimates are not statistically different from one another. Hence, their standard errors will not be statistically different from one another. Therefore, [...] the treatment effects can be pooled by a simple average" (p. 80). Hedges and Olkin (1985) and Rosenthal and Rubin (1986) have proposed two procedures that consider the correlational nature of data. The Rosenthal and Rubin procedure requires the assumption of homogeneous correlations among the p d estimates and is obtained by (Rosenthal & Rubin, 1986, p. 402):[2]

$$d_{RR} = \frac{\sum \lambda_i \, d_i}{\sqrt{\rho \, (\sum \lambda_i)^2 + (1 - \rho)\sum \lambda_i^2}}, \qquad (8)$$

where λ_i is the weight we wish to assign, according to its importance, to the ith dependent variable (or effect size), and ρ is the average intercorrelation among the dependent variables. On the other hand, Hedges and Olkin's proposal (Hedges & Olkin, 1985, p. 212; see also Gleser & Olkin, 1994, p. 352) implies computing an average d value, d_{HO}, taking into account the correlation matrix of the d estimates, via:

$$d_{HO} = \left[\frac{\mathbf{e'} \, \mathbf{S}^{-1}}{\mathbf{e'} \, \mathbf{S}^{-1}\mathbf{e}} \right] \mathbf{d}, \qquad (9)$$

where $\mathbf{e}$ is a p x 1 unitary vector, $\mathbf{d}$ is given by (3), and $\mathbf{S}$ is given by (4). The HO procedure requires the homogeneous effect sizes assumption and they propose testing it by a Chi-square test (Hedges & Olkin, 1985, Eq. 7, p. 211).

The RR and HO procedures have different purposes and estimate different quantities. The RR procedure provides the effect size for a single composite variable made up of the multiple effect sizes. It is analogous, for example, to computing the effect size on the total score obtained on a 25-item test. In terms of correlation coefficient, the RR proposal is equivalent to computing the squared multiple correlation coefficient resulting from regressing the 25 items onto the group membership. Instead, the HO proposal would imply computing 25 individual effect size estimates, one for each item, and estimating their typical value. The HO procedure is a generalization of averaging a set of independent ESs weighted by its inverse-covariance matrix, where the covariances are not null. If the ESs were independent, then the covariance matrix would become a diagonal-variance matrix with zeros as diagonal-off elements. Thus, the HO

---

[2] The equation (1) in Rosenthal and Rubin (1986, p. 402) is a function of the statistical t test for comparing two independent groups:

$$d_{RR} = \frac{\sum \lambda_i t_i / \sqrt{(n - 1)/2}}{\sqrt{\rho(\sum \lambda_i)^2 + (1 - \rho)\sum \lambda_i^2}}, \qquad (12)$$

where $n = n^E = n^C$. Our equation (9) is the same, but rearranged as a function of standardized mean difference, d. To apply this procedure, we give equal weight to the p variables so that $\lambda_1 = \lambda_2 = ... = \lambda_p = 1$.

Table 1
Manipulated Conditions in ES Vectors

ES homogeneity	Effect size magnitude		
	Low	Medium	High
	0.2	0.5	0.8
Maximum	0.2	0.5	0.8
	0.2	0.5	0.8
	0.0	0.3	0.6
Medium	0.2	0.5	0.8
	0.4	0.7	1.0
	–0.4	–0.1	0.2
Low	0.2	0.5	0.8
	0.8	1.2	1.4

procedure is conceptually similar to obtaining a simple mean, however taking into account the correlational structure of the data. In sum, the *RR* procedure estimates a composite ES resulting from pooling all of the dependent variables, achieving more reliability than each individual variable. On the other hand, the *HO* procedure estimates the ES representative of each variable, achieving more efficiency than if only one variable were used.

A comparative study

As mentioned above, the arithmetic mean is an incorrect procedure for averaging dependent ESs, because it ignores the correlational structure of the data, although it is advisable in some conditions. The *RR* and *HO* procedures take into account the intercorrelations among variables, but estimating different parameters; thus, they lead to different quantities; furthermore, factors such as the magnitude and homogeneity of both the effect size vector and the correlation matrix of dependent variables will affect the procedures to a different extent.

In order to show the divergences among the three procedures, a single study with two groups (experimental and control), the same sample size per group ($n^E = n^C = 15$), and three dependent variables ($p = 3$) was defined, and a series of conditions that could affect the results was manipulated (see Tables 1 and 2):

(a) Magnitude of the ESs: vectors of three ESs averaging 0.2, 0.5, and 0.8 were defined, following the criteria in Cohen (1988).

(b) Homogeneity of ESs: three homogeneity levels for *d* estimates were defined, maximum ($\sigma^2 = 0.0$), medium ($\sigma^2 = 0.03$), and low ($\sigma^2 = 0.24$).

(c) Magnitude of the correlation matrix among dependent variables: three conditions were defined in order to obtain mean correlations of .4, .6, and .8.

(d) Homogeneity of the correlation matrix for the dependent variables: two conditions were defined, one with identical correlations, and the other with medium homogeneity.

Thus, a total of 3 (magnitude of ESs) x 3 (homogeneity of ESs) x 3 (magnitude of correlation matrix) x 2 (homogeneity of correlation matrix) = 54 conditions were defined. In each of these conditions, the three procedures for averaging were applied, using the GAUSS (1992) program.

Tables 3 and 4 show the estimates for each procedure as a function of the manipulated conditions.[3] As expected, when

Table 2
Manipulated Conditions in Correlation Matrix

Correlation Matrix Homogeneity	Magnitude of the Correlation Matrix								
	Low			Medium			Maximum		
	1			1			1		
Maximum	0.4	1		0.6	1		0.8	1	
	0.4	0.4	1	0.6	0.6	1	0.8	0.8	1
	1			1			1		
Medium	0.35	1		0.55	1		0.75	1	
	0.4	0.45	1	0.6	0.65	1	0.8	0.85	1

---

[3] The statistically adequate application of the averaging procedures requires the homogeneity assumption of the dependent ESs. After applying the Chi-square test for the homogeneity of dependent ESs (Hedges & Olkin, 1985, p. 211), we found a significant heterogeneity in the conditions labeled as being of low homogeneity ESs ($p < .05$), and a nonsignificant heterogeneity in the other conditions. Although it is inappropriate to apply the procedures in the heterogeneous case, we have included it as a guideline, referring briefly to its trend. On the other hand, it is not usual for meta-analytic studies to test the homogeneity assumption in order to average dependent ESs. Consequently, meta-analyses averaging heterogeneous dependent ESs in their primary studies are very common, and our findings can contribute to assessing the influence of this practice.

Table 3
Estimates in the Homogeneous Correlation Case

ES H	ES M	High ($\bar{r}$ = 0.8)			Medium ($\bar{r}$ = 0.6)			Low ($\bar{r}$ = 0.4)		
		d_M	d_{RR}	d_{HO}	d_M	d_{RR}	d_{HO}	d_M	d_{RR}	d_{HO}
Max.	High	0.800	0.859	0.800	0.800	0.934	0.800	0.800	1.033	0.800
	Med.	0.500	0.537	0.500	0.500	0.584	0.500	0.500	0.645	0.500
	Low	0.200	0.215	0.200	0.200	0.233	0.200	0.200	0.258	0.200
Med.	High	0.800	0.859	0.766	0.800	0.934	0.784	0.800	1.033	0.790
	Med.	0.500	0.537	0.477	0.500	0.584	0.490	0.500	0.645	0.494
	Low	0.200	0.215	0.191	0.200	0.233	0.196	0.200	0.258	0.197
Low	High	0.800	0.859	0.512	0.800	0.934	0.664	0.800	1.033	0.716
	Med.	0.500	0.537	0.310	0.500	0.584	0.411	0.500	0.645	0.445
	Low	0.200	0.215	0.122	0.200	0.233	0.164	0.200	0.258	0.178

Note. ES H = Effect size homogeneity; ES M = Effect size magnitude.

the ES vector was homogeneous (e.g., **d'** = [0.5, 0.5, 0.5]), the *HO* average was equivalent to the individual ES estimates; that is, it estimated the ES of the typical dependent variable of the study, and coincided with a simple average. Nevertheless, the *RR* procedure presented a systematically higher estimate than the constant magnitude, because it estimated a composite ES, adding the nonredundant information in the individual ESs. For example, **d'** = [0.5, 0.5, 0.5] being the ES vector (maximum ES homogeneity and medium ES magnitude), and $\bar{r}$ = 0.6 being the common correlation between the ESs in the homogeneous correlation case, Table 3 shows that $d_{HO} = d_M = 0.5$, whereas the *RR* procedure yielded the higher estimate $d_{RR} = 0.584$. Only

with a perfect correlation among ESs, will the *RR* procedure coincide with the individual estimates, as well as with the *HO* procedure and with a simple average. Following with the analogy of the items and the test, the *RR* procedure represented the overall ES estimate of the test, whereas the *HO* procedure gave an ES estimate of the standard item.

With heterogeneous ESs, the *HO* average was always smaller than the simple mean because of the fact that it is a sum of ESs, each corrected for its regression on the other individual effect sizes. The *RR* procedure, as a composite estimate, presented the highest magnitude. For example, Table 3 shows that $d_{HO} = 0.716$ was lower than $d_M = 0.800$, and so, the *RR* procedure yielded the highest estimate d_{RR}

Table 4
Estimates in the Homogeneous Correlation Case

ES H	ES M	High ($\bar{r}$ = 0.8)			Medium ($\bar{r}$ = 0.6)			Low ($\bar{r}$ = 0.4)		
		d_M	d_{RR}	d_{HO}	d_M	d_{RR}	d_{HO}	d_M	d_{RR}	d_{HO}
Max.	High	0.800	0.859	0.800	0.800	0.934	0.800	0.800	1.033	0.800
	Med.	0.500	0.537	0.500	0.500	0.584	0.500	0.500	0.645	0.500
	Low	0.200	0.215	0.200	0.200	0.233	0.200	0.200	0.258	0.200
Med.	High	0.800	0.859	0.727	0.800	0.934	0.767	0.800	1.033	0.779
	Med.	0.500	0.537	0.439	0.500	0.584	0.472	0.500	0.645	0.482
	Low	0.200	0.215	0.153	0.200	0.233	0.178	0.200	0.258	0.186
Low	High	0.800	0.859	0.381	0.800	0.934	0.611	0.800	1.033	0.682
	Med.	0.500	0.537	0.176	0.500	0.584	0.357	0.500	0.645	0.411
	Low	0.200	0.215	–0.011	0.200	0.233	0.109	0.200	0.258	0.143

Note. ES H = Effect size homogeneity; ES M = Effect size magnitude.

= 1.033 for **d'** = [0.2, 0.8, 1.4] (low ES homogeneity and high ES magnitude), and $\bar{r}$ = 0.4. In any case, the simple mean is a mathematically incorrect procedure of estimation, because of the dependence among ESs. It is only a good approximation when the individual ESs are homogeneous.

Another interesting result was the effect of manipulating the correlation matrix magnitude on the averaging procedures. The lower the correlation among the dependent ESs, the higher the nonredundant information. Thus, both the *RR* and the *HO* estimates decreased their magnitude as the correlation matrix magnitude increased. However, the trend in *RR* was more pronounced than in *HO*, and *HO* was not affected when the ESs vector was homogeneous. For example, Table 3 shows that, for **d'** = [0.6, 0.8, 1], d_{RR} = 1.033 and d_{HO} = 0.790 when $\bar{r}$ = 0.4, whereas d_{RR} = 0.859 and d_{HO} = 0.766 when $\bar{r}$ = 0.8.

Although we have reviewed the performance of the procedures when applied in a single study, in practice, their differences can be accumulated throughout the primary studies in a meta-analysis, greatly affecting the overall results of such a meta-analysis. Furthermore, for greater generalizability of these results, one would require a more comprehensive simulation study, where the meta-analytic conditions usually found in practice (number of studies, sample size of the studies, number of dependent ESs by study, and population ES distribution) would be taken into account. In any case, the numerical discrepancies among the procedures in the example clearly show that it is necessary to clarify the criteria supporting a particular choice among the procedures.

Discussion

In this paper, we have shown the numerical and conceptual differences among three procedures for averaging dependent ESs. The application of the procedures to the same data offers different results, because they estimate different quantities. Thus, the meta-analyst must know the purpose of each procedure, in order to select the one best fitted to his or her aims.

The key difference between the *RR* and the *HO* procedures is in the management of the available information. The *HO* procedure is equivalent to selecting the information of a single ES, the standard one, using the information of the other ESs in order to obtain more efficiency in the estimation. The increase in efficiency depends on the magnitude of the correlations among ESs. With dependent variables estimating the same construct, a relatively high correlation is expected. As Hedges and Olkin (1985, p. 222) pointed out, when the correlations are high, little is gained in efficiency; then, little information is lost by selecting any of the individual ESs or calculating the simple arithmetic mean of the ESs. Assuming the ESs are homogeneous, Marascuillo et al. (1988) also proposed computing the simple arithmetic mean of the ESs.

The *RR* procedure takes into account all of the information contained in the dependent variables, calculating a composite outcome measure. The pooled ES is derived from this composite, taking advantage of the nonredundant information in the individual ESs. The higher reliability of the composite ES justifies the higher magnitude of the *RR* estimate with regard to both the *HO* estimate and the simple mean.

From the point of view of a meta-analyst who intends to take advantage of the higher reliability of a composite ES, the *RR* procedure will be chosen. Other factors being equal, the larger the number of dependent ESs, the larger the composite ES. On the other hand, if the meta-analyst wants to control the different number of dependent variables in a set of primary studies, representing each study by its typical ES, he or she would apply the *HO* procedure. Finally, the simple average among ESs, although technically incorrect, produces suitable estimates when the homogeneity of the dependent ESs is assumed, with very similar results to those of the *HO* procedure.

References

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

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

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

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: Sage.

Gleser, L.J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H.M. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339-355). New York: Sage.

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: Sage.

Johnson, B. T., & Eagly, A. H. (in press). Quantitative synthesis of social psychological research. In H.T. Reis & C.M. Judd (Eds.), *The handbook of research methods in social psychology*. London: Cambridge University Press.

Kalaian, H.A., & Raudenbush, S.W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, *1*, 225-235.

Marascuillo, L.A., Busk, P.L., & Serlin, R.C. (1988). Large sample multivariate procedures for comparing and combining effect sizes within a single study. *Journal of Experimental Education*, *58*, 69-85.

Marín-Martínez, F., & Sánchez-Meca, J. (1998). Testing for dichotomous moderators in meta-analysis. *Journal of Experimental Education, 67*, 69-81.

Matt, G.E. (1989). Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychological Bulletin*, *105*, 106-115.

Matt, G.E., & Cook, T.D. (1994). Threats to the validity of research syntheses. In H.M. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 503-520). New York: Sage.

Raudenbush, S.W., Becker, B.J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, *103*, 111-120.

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

Rosenthal, R. (1994). Parametric measures of effect size. In H.M. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Sage.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, *118*, 183-192.

Rosenthal, R., & Rubin, D.B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin*, *99*, 400-406.

Sánchez-Meca, J., & Ato, M. (1989). Meta-análisis: Una alternativa metodológica a las revisiones tradicionales de la investigación. In J. Arnau & H. Carpintero (Eds.), *Tratado de psicología general. I: Historia, teoría y método* (pp. 617-669). Madrid: Alhambra.

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

Sánchez-Meca, J., & Marín-Martínez, F. (1998a). Weighting by inverse-variance or by sample size in meta-analysis: A simulation study. *Educational and Psychological Measurement, 58,* 211-220

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