British Journal of Mathematical and Statistical Psychology (2017), 70, 439–456
© 2017 The British Psychological Society

The British Psychological Society

www.wileyonlinelibrary.com

Analysis of categorical moderators in mixed-effects meta-analysis: Consequences of using pooled versus separate estimates of the residual between-studies variances

María Rubio-Aparicio[1], Julio Sánchez-Meca[1]*, José Antonio López-López[2], Juan Botella[3] and Fulgencio Marín-Martínez[1]

[1]Department of Basic Psychology & Methodology, Faculty of Psychology, University of Murcia, Spain
[2]School of Social and Community Medicine, University of Bristol, UK
[3]Department of Social Psychology & Methodology, Faculty of Psychology, Autonomous University of Madrid, Spain

Subgroup analyses allow us to examine the influence of a categorical moderator on the effect size in meta-analysis. We conducted a simulation study using a dichotomous moderator, and compared the impact of pooled versus separate estimates of the residual between-studies variance on the statistical performance of the $Q_{B(P)}$ and $Q_{B(S)}$ tests for subgroup analyses assuming a mixed-effects model. Our results suggested that similar performance can be expected as long as there are at least 20 studies and these are approximately balanced across categories. Conversely, when subgroups were unbalanced, the practical consequences of having heterogeneous residual between-studies variances were more evident, with both tests leading to the wrong statistical conclusion more often than in the conditions with balanced subgroups. A pooled estimate should be preferred for most scenarios, unless the residual between-studies variances are clearly different and there are enough studies in each category to obtain precise separate estimates.

1. Introduction

Meta-analysis is a form of systematic review that allows the results of a set of primary studies focused on a common topic to be integrated by the application of statistical methods (Borenstein, Hedges, Higgins, & Rothstein, 2009). While primary studies typically use participants as the unit of analysis, in most meta-analyses the unit of analysis is the study. One of the steps in a meta-analysis consists of synthesizing the results of the primary studies using effect sizes, which can then be statistically combined using meta-analytic techniques. One of the main purposes of meta-analysis is to examine whether the individual effect sizes are homogeneous around the average effect size. When there is more heterogeneity than expected from sampling error, the meta-analyst must search for study characteristics that can explain at least part of that variability. The moderators are

*Correspondence should be addressed to Julio Sánchez-Meca, Department of Basic Psychology & Methodology, Faculty of Psychology, Espinardo Campus, University of Murcia, 30100 Murcia, Spain (email: jsmeca@um.es).

DOI:10.1111/bmsp.12092

considered as potential predictor variables and the effect sizes constitute the dependent variable (Borenstein *et al.*, 2009). If the moderator variable is categorical, an analysis of variance, or subgroup analysis, can be formulated, while the continuous moderators are analysed using meta-analytic analogues to regression analysis.

There are two general statistical models for meta-analysis, the fixed-effect and the random-effects models. The fixed-effect model assumes that all included studies in the meta-analysis share a common population effect size, so the only source of variability is due to sampling error in the selection of the participants of each study (Konstantopoulos & Hedges, 2009). By contrast, the random-effects model assumes that the population effect size could vary from study to study due to differential characteristics of the studies. Consequently, this model assumes a distribution of the population effect sizes and adds a second source of variability, the sampling error in the selection of the studies in the meta-analysis (Raudenbush, 2009). Note that the random-effects model assumes the more realistic scenario of heterogeneity among the population effect sizes, due to the differential characteristics of the studies in a meta-analysis.

1.1. Subgroup analysis

In meta-analysis, the analysis of categorical moderators is usually referred to as 'subgroup analysis', and is the process of comparing the mean effect sizes in different study subgroups (Borenstein & Higgins, 2013).

Several statistical models are available to examine the relationship between a categorical moderator and the effect sizes through a subgroup analysis. On the one hand, applying the logic of the general fixed-effect model to subgroup analyses, a fixed-effects model can be assumed in which all studies within the same category of the moderator share a common effect size. In other words, if a fixed-effect model is assumed within each subgroup, such model is called a fixed-effects model.

On the other hand, the mixed-effects model consists of assuming a random-effects model for each subgroup of studies. As a consequence, the mixed-effects model assumes that all studies within the same category of the moderator estimate a normal distribution of population effect sizes with a common mean effect size. The label 'mixed-effects model' is used because: (1) the moderator is considered a fixed-effects component, as the categories of the moderator are not a random sample of a larger number of categories, and (2) the effect sizes (i.e., the studies) include a random-effects component because they are considered a random sample of study effects pertaining to a population of studies in the same category (Borenstein *et al.*, 2009; Viechtbauer, 2010).

In this paper, we focus on the performance of the mixed-effects model, which is nowadays routinely applied in most meta-analytic studies.

1.2. Mixed-effects model

Suppose that the k studies in a meta-analysis are grouped into m mutually exclusive categories of the moderator variable. Denote by $k_1, k_2, \ldots, k_m$ the number of effect sizes of categories 1, 2, …, m, respectively, such that $k_1 + k_2 + \ldots + k_m = k$.

In a mixed-effects model the individual effect sizes, T_{ij}, within the same category j are assumed to estimate a distribution of true effect sizes with mean $\mu_{\theta j}$ and variance $\sigma_{ij}^2 + \tau_j^2$, with σ_{ij}^2 being the within-study variance for the ith study in the jth category of the moderator, and τ_j^2 the residual between-studies variance in that category.

We must assume a random-effects model within each category of the moderator variable, thus the statistical model applied in the *j*th category will be $T_{ij} = \mu_{\theta j} + \varepsilon_{ij} + e_{ij}$, where ε_{ij} and e_{ij} are the within-study and between-studies errors, respectively. It is very common to assume that these two errors are independent of each other and, therefore, the estimated effect sizes are normally distributed: $T_{ij} \sim N(\mu_{\theta j}, \sigma^2_{ij} + \tau^2_j)$, where τ^2_j is the common between-studies variance in *j*th category of the moderator. In addition, the parametric effect sizes of the *j*th category, θ_{ij}, follow a normal distribution with mean $\mu_{\theta j}$ and between-studies variance τ^2_j: $\theta_{ij} \sim N(\mu_{\theta j}, \tau^2_j)$.

Under a mixed-effects model, the main goal in a subgroup analysis is to compare the parametric mean effect sizes from each category of the moderator variable, $\mu_{\theta j}$, in order to test if the moderator is statistically related to the effect sizes. Consequently, first we need to estimate the mean parametric effect size of the *j*th category of the moderator, $\mu_{\theta j}$, by means of

$$\bar{T}_j = \frac{\sum_i \hat{w}_{ij} T_{ij}}{\sum_i \hat{w}_{ij}}, \tag{1}$$

where $\hat{w}_{ij}$ are the estimated weights computed through $\hat{w}_{ij} = 1\Big/\left(\hat{\sigma}^2_{ij} + \hat{\tau}^2_j\right)$, with $\hat{\sigma}^2_{ij}$ being the estimated within-study variance of the *i*th effect size and $\hat{\tau}^2_j$ the estimated residual between-studies variance of the *j*th category.

The sampling variance of the mean effect size in the *j*th category is estimated as

$$V(\overline{T}_j) = \frac{1}{\sum\limits_i \hat{w}_{ij}}. \tag{2}$$

1.3. Omnibus test of between-groups differences

It is possible to test the statistical significance of a categorical moderator by means of the between-groups heterogeneity statistic, given by (Borenstein *et al.*, 2009)

$$Q_{\mathrm{B}} = \sum_{j=1}^{m} \hat{w}_{+j}\left(\overline{T}_j - \overline{T}\right)^2, \tag{3}$$

where $\hat{w}_{+j}$ is the inverse of equation (2) applied to the *j*th category of the moderator, $\bar{T}_j$ is the mean effect size of the *j*th category calculated by equation (1) and $\overline{T}$ represents the weighted grand mean of all effect sizes and is given by

$$\overline{T} = \frac{\sum\limits_i \sum\limits_j \hat{w}_{ij} T_{ij}}{\sum\limits_i \sum\limits_j \hat{w}_{ij}}, \tag{4}$$

where the total between-studies variance estimate, $\hat{\tau}^2$, is used to compute $\hat{w}_{ij}$.

Under the null hypothesis of no difference between the mean effect sizes for each of the *m* categories (H_0: $\mu_{\theta 1} = \mu_{\theta 2} = \ldots = \mu_{\theta m}$), the Q_{B} statistic follows a chi-square distribution with $m - 1$ degrees of freedom. Therefore, the null hypothesis will be rejected when Q_{B} exceeds the 100(1 − α) percentile point of the chi-square distribution. A statistically significant result for Q_{B} provides evidence that the moderator is statistically related to the effect sizes.

1.4. Estimating the residual between-studies variance

Several methods have been proposed to estimate the total heterogeneity variance in the random-effects model. The most commonly used is that proposed by DerSimonian and Laird (1986), a heterogeneity variance estimator derived from the moment method.

At this point, it may be useful to make a distinction between the total between-studies variance and the residual between-studies variance. On the one hand, when we apply the random-effects model to estimate the mean effect in a meta-analysis (i.e., without moderators being added to the model) there is an amount of heterogeneity due to sampling error in the selection of the studies in the meta-analysis. This heterogeneity is estimated through the total between-studies variance, which represents the excess variation among the effects over that expected from within-study sampling error alone. On the other hand, in the mixed-effects model we include moderator variables aiming to explain at least part of the total heterogeneity in the effect sizes. Thus, after adding moderator variables the amount of heterogeneity that remains to be explained is the residual heterogeneity or the heterogeneity that cannot be explained by the moderators included in the model.

In the mixed-effects model, two approaches can be adopted to estimate the residual between-studies variance. One is to estimate the residual between-studies variance separately within each category of the moderator, and the other one is to calculate a pooled estimate across categories (Borenstein *et al.*, 2009).

1.4.1. Separate estimates of the residual between-studies variance

This procedure consists of estimating the residual between-studies variance within each category of the moderator. Thus, in a moderator variable with m categories, we need to calculate the residual between-studies variance estimates $\hat{\tau}_1^2, \hat{\tau}_2^2, \ldots, \hat{\tau}_m^2$. The residual between-studies variance for the jth category of the moderator, $\hat{\tau}_j^2$, can be computed applying the DerSimonian and Laird estimator with the expression

$$\hat{\tau}_j^2 = \frac{Q_{\mathrm{W}j} - (k_j - 1)}{c_j}, \tag{5}$$

where k_j is the number of studies of the jth category, $Q_{\mathrm{W}j}$ is the within-group homogeneity statistic of the jth category computed as

$$Q_{\mathrm{W}j} = \sum_{i=1}^{k_j} \hat{w}_{ij}^* \left(T_{ij} - \bar{T}_j^* \right), \tag{6}$$

with $\hat{w}_{ij}^*$ being the estimated weights assuming a fixed-effect model, $\hat{w}_{ij}^* = 1 \big/ \hat{\sigma}_{ij}^2$, and $\bar{T}_j^*$ the mean effect size of the jth category of the moderator also assuming a fixed-effect model, that is, applying equation (1) but using $\hat{w}_{ij}^*$ as weighting factor; and c_j is given by

$$c_j = \sum_i \hat{w}_{ij}^* - \frac{\sum_i \left(\hat{w}_{ij}^* \right)^2}{\sum_i \hat{w}_{ij}^*}. \tag{7}$$

Therefore, equation (5) allows a separate estimate of the between-studies variance of each category, $\hat{\tau}_j^2$, to be obtained, and these are used to calculate the weights, $\hat{w}_{ij}$, for each category of the moderator. This implies that in each category a different between-studies variance is used to calculate the weights: $\hat{\tau}_1^2$ for category 1, $\hat{\tau}_2^2$ for category 2, and so on, that is, $\hat{w}_{ij} = 1\Big/\left(\hat{\sigma}_{ij}^2 + \hat{\tau}_j^2\right)$. Here we will denote the Q_B statistic calculated with separate between-studies variances by $Q_{\mathrm{B(S)}}$.

1.4.2. Pooled estimate of the residual between-studies variance

An alternative method to estimate the residual heterogeneity variance consists of averaging the residual between-studies variances of the m categories of the moderator variable, through the equation (Borenstein *et al.*, 2009)

$$\hat{\tau}_+^2 = \frac{\sum_j^m Q_{\mathrm{W}j} - \sum_j^m (k_j - 1)}{\sum_j^m c_j}. \tag{8}$$

Equation (8) provides a pooled estimate of the residual between-studies variance, so that the weights, $\hat{w}_{ij}$, are obtained using a common between-studies variance through the different categories of the moderator, that is, $\hat{w}_{ij} = 1\Big/\left(\hat{\sigma}_{ij}^2 + \hat{\tau}_+^2\right)$. Here we will use the term $Q_{\mathrm{B(P)}}$ to refer to the Q_B statistic calculated with a pooled estimate of the residual between-studies variance, $\hat{\tau}_+^2$.

1.5. An example

To illustrate how the Q_B statistic is calculated with the two different methods to estimate the residual between-studies variance (pooled vs. separate estimates), an example extracted from a real meta-analysis is presented here. The data were obtained from a meta-analysis of the efficacy of psychological treatments for panic disorder with or without agoraphobia (Sánchez-Meca, Rosa-Alcázar, Marín-Martínez, & Gómez-Conesa, 2010). The effect size index in this meta-analysis was the standardized mean difference (d) between two groups (treated vs. control groups) defined in equation (10) below. From all the moderator variables analysed in this meta-analysis, a dichotomous characteristic was selected to illustrate a subgroup meta-analysis: whether or not the assignment of the participants to the treated and control groups was at random. The database composed of 50 studies is presented in Appendix.

Tables 1 and 2 present the results yielded by the Q_B statistic with the two methods here compared, as well as the mean effects for each category of the moderator, the sampling variances, the residual between-studies variances and the 95% confidence intervals for each mean effect. Separate estimates of the residual between-studies variances for each category ($\hat{\tau}_j^2$) were calculated using equation (5). As shown in Table 1, their values were 0.053 and 0.303 for non-random and random assignment, respectively. On the other hand, the pooled estimate of the residual between-studies variances calculated using equation (8) was $\hat{\tau}_+^2$ = 0.270 (Table 2). When the Q_B statistic was calculated taking separate estimates of the residual between-studies variances, the estimated weights for each study were obtained by means of $\hat{w}_{ij} = 1\big/(\hat{\sigma}_{ij}^2 + \hat{\tau}_j^2)$. Conversely, when the Q_B statistic was calculated taking a pooled estimate of the residual between-studies variances ($\hat{\tau}_+^2$) the estimated study weights were $\hat{w}_{ij} = 1\big/(\hat{\sigma}_{ij}^2 + \hat{\tau}_+^2)$. This distinction affects the Q_B statistic, here denoted by $Q_{\mathrm{B(S)}}$ and $Q_{\mathrm{B(P)}}$, respectively, as well as the mean effect from each category of the moderator, their sampling variances ($V(\overline{d}_j)$), and their confidence limits.

Table 1. Results of the subgroup analysis for the moderator variable 'random assignment' in the Sánchez-Meca *et al.* (2010) meta-analysis by using separate estimates of the residual between-studies variance, $\hat{\tau}_j^2$

				95% CI		
Random assignment	k_j	$\bar{d}_j$	$V(\bar{d}_j)$	d_1	d_2	$\hat{\tau}_j^2$
No	8	0.545	0.024	0.242	0.847	0.053
Yes	42	0.966	0.011	0.765	1.167	0.303
Separate estimates of $\hat{\tau}_j^2$:	$Q_{B(S)}(1) = 5.165$, $p = .023$					

Notes. k_j = number of studies in each category of the moderator. $\bar{d}_j$ = mean effect size for each category, obtained with equation (1). $V(\bar{d}_j)$= estimated sampling variance of the mean effect size for each category, obtained with equation (2). d_L and d_U = lower and upper confidence limits (for a 95% confidence level) for each mean effect size, obtained from $\bar{d}_j \pm 1.96 \times \sqrt{V(\bar{d}_j)}$ (1.96 being the 97.5 percentile of the standard normal distribution). $\hat{\tau}_j^2$ = residual between-studies variance for each category, estimated with equation (5).

The mean effects for non-random and random assignment were 0.545 and 0.966, respectively (Table 1), when separate estimates of the residual between-studies variances were used ($\hat{\tau}_j^2$), and 0.559 and 0.961 when a pooled estimate ($\hat{\tau}_+^2$) was used (Table 2). The sampling variances and the confidence limits also varied depending on the residual between-studies variances used in the calculations. However, the most dramatic discrepancy among methods involved the two versions of the Q_B statistic: $Q_{B(S)}$ and $Q_{B(P)}$. The null hypothesis of equal mean effect sizes was rejected when separate estimates of the between-studies variances were used (Table 1: $Q_{B(S)} = 5.165$, $p = .023$), but not when a pooled estimate was considered (Table 2: $Q_{B(P)} = 2.588$, $p = .108$).

This example illustrates how results and their interpretation can be affected by the meta-analytic methods selected to undertake the statistical analyses. The choice of the meta-analyst will often be conditioned by the software used for the calculations and he/she will not be aware of which method was implemented. In fact, the most commonly used statistical programs for meta-analysis do not enable users to choose among the two methods to calculate the individual weights in a mixed-effects model. For instance, if the meta-analyst used metafor (Viechtbauer, 2010), Comprehensive Meta-Analysis 2.0 (Borenstein, Hedges, Higgins, & Rothstein, 2005) or the SPSS macros written by David B. Wilson to replicate this example, he/she would obtain the results presented in Table 2, whereas if using RevMan 5.3 (Review Manager, 2014), the results would be those presented in Table 1. On the other hand, Comprehensive Meta-Analysis 3.0 (Borenstein, Hedges, Higgins, & Rothstein, 2014) incorporates both methods so that the meta-analyst can use either to estimate the weights (in fact, the results in Tables 1 and 2 were obtained with this program).

1.6. Purpose of the study

It is not clear which of these two procedures (separate or pooled estimates) should be preferred in order to estimate the residual between-studies variance, which is involved in the subgroup analysis in a mixed-effects meta-analysis. At this point, it is useful to revise the analogy between the subgroup analysis in meta-analysis and the analysis of variance (ANOVA) for comparing means in a primary study. On the one hand, in the simplest case

Table 2. Results of the subgroup analysis for the moderator variable 'random assignment' in the Sánchez-Meca *et al.* (2010) meta-analysis by using a pooled estimate of the residual between-studies variance, $\hat{\tau}^2_+$

				95% CI		
Random assignment	k_j	$\bar{d}_j$	$V(\bar{d}_j)$	d_L	d_U	$\hat{\tau}^2_+$
No	8	0.559	0.053	0.109	1.009	0.270
Yes	42	0.961	0.010	0.768	1.155	0.270
Pooled estimate of $\hat{\tau}^2_j$:	$Q_{B(P)}(1) = 2.588, p = .108$					

Notes. k_j = number of studies in each category of the moderator. $\bar{d}_j$ = mean effect size for each category, obtained with equation (1). $V(\bar{d}_j)$ = estimated sampling variance of the mean effect size for each category, obtained with equation (2). d_L and d_U = lower and upper confidence limits (for a 95% confidence level) for each mean effect size, obtained by means of $\bar{d}_j \pm 1.96 \times \sqrt{V(\bar{d}_j)}$ (1.96 being the 97.5 percentile of the standard normal distribution). $\hat{\tau}^2_+$ = pooled estimate of the residual between-studies variances of the two categories, calculated with equation (8).

of a primary study with a two-independent-group design (e.g. experimental vs. control groups), the means of two samples of subjects are compared performing a *t*-test or an ordinary least squares ANOVA. On the other hand, in a meta-analysis with two subgroups of studies, the mean effect sizes in each subgroup are compared by performing a weighted least squares ANOVA, the weights being the inverse variance of each effect size.

Both the *t*-test and ANOVA for comparing the means of two or more independent groups of subjects assume homogeneity between variances in the two populations. The pooled variance is estimated through the mean squared error in the ANOVA. When the two population variances are heterogeneous, the so-called Behrens–Fisher problem arises, which requires an alternative procedure to the classical *t*-test or ANOVA. In practice, the usual solution to the Behrens–Fisher problem is to apply the Welch–Satterthwaite approach to correct the classical *t*-test (Welch, 1947).

In the meta-analytic arena, the picture is a little more complex, as we are working with aggregate scores (e.g. effect sizes summarizing individual scores) instead of individual participants. While in a primary study each subject provides a score, in a meta-analysis each study provides an effect size. The effect sizes of the studies in a meta-analysis will exhibit different precision depending of the sample size of the study. Effect sizes obtained from large samples will be more accurate (less variable) than those obtained from small ones. As a consequence, the appropriate mean of a set of effect sizes is a weighted average, the weights being the inverse variance of each effect size. This weighting scheme affects all statistical calculations in a meta-analysis.

The pooled estimation of the residual between-studies variance from two or more subgroups of studies in a meta-analysis is akin to the estimation of the mean squared error in the ANOVA in a primary study, as both procedures assume the variance between groups to be homogeneous. When this assumption is not tenable, a similar problem to that of Behrens–Fisher emerges, which may lead to inaccurate estimation of the residual between-studies variance. To circumvent this problem, an alternative is the separate estimation of the residual between-studies variance for each subgroup of studies. However, this approach can also yield inaccurate estimates if the number of studies in the subgroups is small (which will often be the case).

In a mixed-effects meta-analysis, the residual between-studies variance is included in the weighting scheme. Thus, the estimation procedure for the residual between-studies variance may have an impact on a wide range of meta-analytic outputs, such as: (1) the estimate of the average effect size for each category of the moderator (see equation (1)); (2) their sampling variances; (3) the confidence intervals; and, relevant to the present work, (4) the computation of the between-groups heterogeneity statistic, Q_{B} (see equation (3)).

The large number of factors that can affect the performance of the $Q_{\mathrm{B(P)}}$ and $Q_{\mathrm{B(S)}}$ statistics lead to the need for simulation studies to determine which of them is a better option under different meta-analytic conditions. Previous simulation studies have examined the statistical performance of the *t*-test and ANOVA *F* test in a primary study, assuming homogeneous and heterogeneous population variances. However, those studies do not address the more complex picture of subgroup analyses in meta-analysis, and therefore their findings might not be generalizable to the meta-analytic arena.

The purpose of this work was to directly compare, by means of Monte Carlo simulation, the statistical performance of the Q_{B} statistic applied in meta-analysis, when two alternative procedures for estimating the residual between-studies variance (separate estimates and pooled estimate) are used. With that aim, the present work is the first simulation study to compare the $Q_{\mathrm{B(S)}}$ and $Q_{\mathrm{B(P)}}$ tests, assessing their Type I error and statistical power in different meta-analytic scenarios.

The existence of previous simulation studies addressing the heteroscedasticity problem in primary studies enables us to formulate some expectations (Glass & Hopkins, 1996; Glass, Peckham, & Sanders, 1972; Hinkle, Wiersma, & Jurs, 2003; Senn, 2008). First, in scenarios with balanced sample sizes, we expect $Q_{\mathrm{B(P)}}$ to provide an adequate adjustment of the Type I error, even with heterogeneous variances between subgroups. Second, in unbalanced scenarios with heterogeneous variances where the larger variance is associated with the bigger subgroup, the $Q_{\mathrm{B(P)}}$ test will be too conservative, and too liberal if the smaller variance is associated with the bigger subgroup instead.

2. Method of the simulation study

A simulation study was carried out in R using the metafor package (Viechtbauer, 2010) and the two procedures (pooled and separate) for estimating the residual between-studies variance were programmed. Meta-analyses of *k* studies were simulated with the standardized mean difference as the effect size index. Each individual study included in a meta-analysis compared two groups (experimental and control) with respect to some continuous outcome. Both populations were normally distributed with homogeneous variances ($N(\mu_{\mathrm{E}}, \sigma^2)$, $N(\mu_{\mathrm{C}}, \sigma^2)$). The population standardized mean difference, δ, was defined as (Hedges & Olkin, 1985)

$$\delta = \frac{\mu_{\mathrm{E}} - \mu_{\mathrm{C}}}{\sigma}. \tag{9}$$

The parametric effect size, δ, can be estimated by means of

$$d = c(m)\frac{\bar{y}_E - \bar{y}_C}{S}, \tag{10}$$

where $\bar{y}_E$ and $\bar{y}_C$ are the sample means of experimental and control groups, S is a pooled standard deviation computed as

$$S = \sqrt{\frac{(n_E - 1)S_E^2 + (n_C - 1)S_C^2}{n_E + n_C - 2}} \tag{11}$$

n_E and n_C being the experimental and control sample sizes, respectively, S_E^2 and S_C^2 being the unbiased variances of the two groups, and $c(m)$ is a correction factor for small sample sizes, given by

$$c(m) = 1 - \frac{3}{4N - 9}, \tag{12}$$

with $N = n_E + n_C$. The estimated within-study variance of d, assuming equal variances and normality within each study, is given by

$$\hat{\sigma}_d^2 = \frac{n_E + n_C}{n_E n_C} + \frac{d^2}{2(n_E + n_C)}. \tag{13}$$

We simulated a mixed-effects model involving a moderator variable with two categories. In each category of the moderator variable a population of parametric effect sizes was assumed, in addition to the within-group variability.

The number of studies of each simulated meta-analysis was defined as $k = k_1 + k_2$, with k_1 and k_2 being the number of studies falling into the first and second categories of the moderator, respectively.

The manipulated conditions in the present study were intended to represent the most realistic scenarios found in meta-analysis. For the number of studies, k, we considered four values, 12, 20, 40, and 60. Furthermore, we manipulated the distribution of k within each category of the moderator, so that in some conditions there was a balanced distribution (e.g. $k_1 = k_2$), while in the remaining conditions there was an unbalanced distribution between the two categories with the second category containing three times as many studies as the first category.

We also manipulated the residual between-studies variance of each category of the moderator in two different ways. First, we considered two values for this parameter, namely 0.08 and 0.16. Second, we simulated a set of scenarios with homogeneous residual between-studies variances for both categories ($\tau_1^2 = \tau_2^2$), and also another set of heterogeneous conditions, with values $\tau_1^2 = 0.08$ and $\tau_2^2 = 0.16$ or $\tau_1^2 = 0.16$ and $\tau_2^2 = 0.08$.

The average sample size of the k studies in a meta-analysis was set to 60. Note that, for each study, $N = n_E + n_C$, with $n_E = n_C$. The selection of the sample sizes for the individual studies in each meta-analysis was performed from the generation of skewed distributions, applying the Fleishman's (1978) algorithm with an average value of 60, a skewness index of 1.386, a kurtosis index of 1.427 and a standard deviation of 5.62. The parameters of this distribution are similar to the distribution of sample sizes found in a recent review of 50 real meta-analyses on the effectiveness of psychological treatments (López-López, Rubio-Aparicio, Sánchez-Meca, & Marín-Martínez, 2013).

The parametric mean effect size of each category of the moderator was also manipulated. In some conditions the two parametric mean effects were equal to 0.5 ($\mu_{\delta 1} = \mu_{\delta 2} = 0.5$), whereas for other conditions they were set to different values: $\mu_{\delta 1} = 0.5$ and $\mu_{\delta 2} = 0.3$ or $\mu_{\delta 1} = 0.5$ and $\mu_{\delta 2} = 0.1$. Moreover, when the parametric mean effect sizes were different for each category, their position was also manipulated, and hence we also generated scenarios with $\mu_{\delta 1} = 0.3$ and $\mu_{\delta 2} = 0.5$ or $\mu_{\delta 1} = 0.1$ and

$\mu_{\delta 2} = 0.5$. The conditions with equal parametric mean effect sizes across categories allowed us to study the Type I error rate of the $Q_{B(S)}$ and $Q_{B(P)}$ statistics, whereas the conditions with different parametric mean effect sizes enabled us to assess their statistical power.

To assess the Type I error rate, the total number of conditions was: 4 (number of studies) × 2 (balanced–unbalanced number of studies in the two categories) × 4 (residual between-studies variance) = 32. With respect to the statistical power, the conditions were quadrupled regarding those of the Type I error by including two different parametric mean effect sizes and manipulating their position across categories, so that there were 32 × 4 = 128 conditions defined. To sum up, the total number of conditions was 160, and for each one 10,000 replications were generated. Thus, 1,600,000 meta-analyses were simulated.

The $Q_{B(S)}$ test (equation (3)) using separate estimates of τ^2 for each subgroup (equation (5)) and the $Q_{B(P)}$ test using a pooled estimate of τ^2 (equation (8)) were applied to each one of these replications. In each of the 160 conditions of our simulation study, the proportion of rejections of the null hypothesis of equality of the parametric mean effect sizes of the moderator enabled us to estimate the Type I error rate and the statistical power.

3. Results

3.1. Type I error rate

Table 3 presents Type I error rates for the $Q_{B(S)}$ and $Q_{B(P)}$ statistics when using the two estimation procedures of the residual between-studies variance in the manipulated conditions. Table 4 summarizes the average Type I error rates as a function of the number of studies, balanced and unbalanced distribution of number of studies within each category of the moderator, and residual between-studies variance of each category of the moderator. Note that the nominal significance level was set to $\alpha = .05$.

First, in most conditions results showed the empirical rejection rates of both estimation procedures above the nominal significance level (Tables 3 and 4). As expected, as the number of studies increased, the proportion of rejections of the null hypothesis of equality for $Q_{B(S)}$ and $Q_{B(P)}$ converged to the nominal significance level (Table 4).

In general, when the number of studies was balanced across categories, both estimation procedures showed a good adjustment to the nominal level, with negligible differences among the empirical error rates. By contrast, under the conditions with an unbalanced distribution of studies between the two categories, the differences in error rates for both estimation procedures were most notable (Table 3).

As can be seen in Table 3, and focusing on unbalanced distribution of the number of studies within each category of the moderator, when the residual between-studies variances of each category were homogeneous ($\tau_1^2 = \tau_2^2 = 0.08$ or $\tau_1^2 = \tau_2^2 = 0.16$), the $Q_{B(P)}$ test showed better control of the Type I error rate than the $Q_{B(S)}$ test. In contrast, when variances were heterogeneous, specifically under the condition where the value of the smallest residual between-studies variance, $\tau^2 = 0.08$, was associated with the category with the smallest number of studies ($\tau_1^2 = 0.08$, $\tau_2^2 = 0.16$), the $Q_{B(P)}$ test showed Type I error rates below 0.05, whereas the $Q_{B(S)}$ test yielded rates over nominal except for a large number of studies, $k = 60$ ($k_1 = 15$ and $k_2 = 45$). Under the condition where the value of the largest residual between-studies variance, $\tau^2 = 0.16$, was associated with the category with the smallest number of studies, ($\tau_1^2 = 0.16$, $\tau_2^2 = 0.08$),

Table 3. Type I error for the two estimation procedures of the residual between-studies variance

		Balanced		Unbalanced	
$\tau_1^2 : \tau_2^2$	k	$Q_{B(S)}$	$Q_{B(P)}$	$Q_{B(S)}$	$Q_{B(P)}$
0.08: 0.08	12	0.0611	0.0655	0.0801	0.0719
	20	0.0595	0.0609	0.0743	0.0672
	40	0.0584	0.0581	0.0639	0.0577
	60	0.0543	0.0548	0.0564	0.0527
0.16: 0.16	12	0.0737	0.0761	0.0950	0.0976
	20	0.0648	0.0650	0.0783	0.0652
	40	0.0554	0.0548	0.0696	0.0612
	60	0.0567	0.0566	0.0640	0.0579
0.08: 0.16	12	0.0705	0.0733	0.0758	0.0524
	20	0.0602	0.0611	0.0709	0.0456
	40	0.0584	0.0580	0.0623	0.0377
	60	0.0510	0.0505	0.0552	0.0349
0.16: 0.08	12			0.0956	0.1013
	20			0.0886	0.0949
	40			0.0716	0.0890
	60			0.0606	0.0801

Notes. τ_1^2 = residual between-studies variance of the first category of the moderator; τ_2^2 = residual between-studies variance of the second category of the moderator; k = number of studies; Balanced = balanced distribution of k within each category of the moderator; Unbalanced = unbalanced distribution of k within each category of the moderator, with fewer studies in the first category; $Q_{B(S)} = Q_B$ test using separate estimates of τ^2 for each subgroup; $Q_{B(P)} = Q_B$ test using a pooled estimate of τ^2.

Table 4. Average Type I rates by number of studies (k), by balanced and unbalanced distribution of k, and by the residual between-studies variance of each category of the moderator ($\tau_1^2 : \tau_2^2$)

	$Q_{B(S)}$	$Q_{B(P)}$
K		
12	0.0788	0.0738
20	0.0709	0.0657
40	0.0628	0.0595
60	0.0569	0.0553
Distribution of k		
Balanced	0.0577	0.0577
Unbalanced	0.0679	0.0620
$\tau_1^2 : \tau_2^2$		
0.08: 0.08	0.0612	0.0585
0.16: 0.16	0.0648	0.0601
0.08: 0.16	0.0597	0.0479
0.16: 0.08	0.0736	0.0880

Note. $Q_{B(S)} = Q_B$ test using separate estimates of τ^2 for each subgroup; $Q_{B(P)} = Q_B$ test using a pooled estimate of τ^2.

the $Q_{B(P)}$ test showed empirical rejection rates above the nominal significance level, while the $Q_{B(S)}$ test only showed results close to the nominal level with $k = 60$ ($k_1 = 15$ and $k_2 = 45$).

3.2. Statistical power

Table 5 shows the empirical power rates for $Q_{B(S)}$ and $Q_{B(P)}$ tests in the manipulated conditions. Table 6 summarizes the average power rates as a function of the magnitude of the difference between the parametric mean effect sizes of each category of the moderator, number of studies, balanced and unbalanced distribution of number of studies within each category of the moderator, and the residual between-studies variance for each category of the moderator. In general, the influence of the different conditions manipulated was equivalent for the $Q_{B(S)}$ and $Q_{B(P)}$ tests and, in most conditions, both tests yielded statistical power rates far below .80 (Tables 5 and 6).

Table 6 shows that, as expected, $Q_{B(S)}$ and $Q_{B(P)}$ tests increased their statistical power as the number of studies and the magnitude of the difference between the parametric effect size of each category increased. Furthermore, under the conditions with a balanced distribution of the studies across categories, the $Q_{B(S)}$ and $Q_{B(P)}$ tests showed greater power than under the condition with an unbalanced distribution of the studies (see also Table 5). In relation to the conditions with homogeneous residual between-studies variances, large amounts of residual τ^2 values correspond to smaller rejection rates for both tests. Accordingly, the highest power rates, $Q_{B(S)} = 0.9760$ and $Q_{B(P)} = 0.9759$, were obtained under optimal scenarios, that is, maximum difference between the parametric mean effect size of each category ($|\mu_{\delta 1} - \mu_{\delta 2}| = 0.4$), large number of studies ($k = 60$), balanced distribution of studies within each category and small and homogeneous values of the residual between-studies variance of each category ($\tau_1^2 = 0.08$, $\tau_2^2 = 0.08$) (Table 5).

As shown in Table 5, under a balanced distribution of the number of studies within each category of the moderator, the $Q_{B(S)}$ and $Q_{B(P)}$ tests performed very similarly, even when the assumption of homogeneity variances was not fulfilled. By contrast, when the number of studies was distributed unequally within each category of the moderator and the residual between-studies variances of each category were homogeneous, the $Q_{B(S)}$ test yielded a slightly higher power than the $Q_{B(P)}$ test.

4. Discussion

This study compares the impact of two procedures for estimating the residual between-studies variance, separate estimates and pooled estimate, on the statistical performance of the Q_B test for subgroup analyses assuming a mixed-effects meta-analysis. Our work is the first simulation study to address the question of which estimation procedure for the residual between-studies variance yields the most accurate results for the Q_B test under a set of realistic scenarios, and also allows us to explore the practical consequences of using separate estimates or a pooled estimate.

Under a balanced distribution of the number of studies across categories, we expected good performance from the $Q_{B(P)}$ test even when the assumption of homogeneity of the residual between-studies variances was not fulfilled. This is a similar situation to that of the typical ANOVA *F* test with equal sample sizes between groups of subjects, where the *F* test is robust to violations of the homoscedasticity assumption (Glass & Hopkins, 1996; Senn, 2008). Our results showed similar Type I error rates for the $Q_{B(P)}$ test in the conditions with homogeneous and heterogeneous residual between-studies variances. However, the empirical Type I error rates showed a good adjustment to the nominal level only in meta-analyses with a large number of studies (40 or more studies), the adjustment becoming slightly more liberal as the number of studies decreased.

Table 5. Statistical power rates for the two estimation procedures of the residual between-studies variance

		$\lvert\mu_{\delta 1}-\mu_{\delta 2}\rvert = 0.2$				$\lvert\mu_{\delta 1}-\mu_{\delta 2}\rvert = 0.4$			
		Balanced		Unbalanced		Balanced		Unbalanced	
$\tau_1^2 : \tau_2^2$	*k*	$Q_{B(S)}$	$Q_{B(P)}$	$Q_{B(S)}$	$Q_{B(P)}$	$Q_{B(S)}$	$Q_{B(P)}$	$Q_{B(S)}$	$Q_{B(P)}$
0.08: 0.08	12	0.161	0.1701	0.1599	0.151	0.4383	0.4479	0.3645	0.3638
	20	0.2203	0.2235	0.1894	0.1827	0.6341	0.6385	0.5293	0.5298
	40	0.3796	0.3783	0.3028	0.2953	0.8988	0.9000	0.8028	0.8068
	60	0.5224	0.5220	0.4168	0.4116	0.9760	0.9759	0.9296	0.9323
0.16: 0.16	12	0.1446	0.1483	0.1505	0.1294	0.3298	0.3329	0.3012	0.2792
	20	0.1752	0.1768	0.1642	0.1489	0.4803	0.4804	0.4004	0.3893
	40	0.2756	0.2753	0.2269	0.2175	0.7501	0.7502	0.6305	0.6285
	60	0.3710	0.3700	0.3139	0.3060	0.8979	0.8971	0.7972	0.7994
0.08: 0.16	12	0.1512	0.1567	0.1405	0.1046	0.3759	0.3831	0.3342	0.2635
	20	0.1986	0.2025	0.1749	0.1261	0.5392	0.5443	0.4772	0.4022
	40	0.3136	0.3198	0.2802	0.2130	0.8275	0.8299	0.7542	0.6905
	60	0.4377	0.4432	0.3787	0.3024	0.9478	0.9493	0.9007	0.8615
0.16: 0.08	12	0.1466	0.1512	0.3808	0.1749	0.3677	0.3729	0.3204	0.3541
	20	0.1918	0.1922	0.1778	0.2062	0.5441	0.5443	0.4271	0.4823
	40	0.3146	0.3098	0.2489	0.2960	0.8241	0.8213	0.6763	0.7373
	60	0.4355	0.4274	0.3249	0.3832	0.9432	0.9422	0.8268	0.8748

Notes. $\mu_{\delta 1}$ = parametric mean effect size of the first category of the moderator; $\mu_{\delta 2}$ = parametric mean effect size of the second category of the moderator; τ_1^2 = residual between-studies variance of the first category of the moderator; τ_2^2 = residual between-studies variance of the second category of the moderator; *k* = number of studies; Balanced = balanced distribution of *k* within each category of the moderator; Unbalanced = unbalanced distribution of *k* within each category of the moderator, where the number of studies in the first category is the lowest; $Q_{B(S)}$ = Q_B test using separate estimates of τ^2 for each subgroup; $Q_{B(P)}$ = Q_B test using a pooled estimate of τ^2.

Comparing the performance of the $Q_{B(S)}$ and $Q_{B(P)}$ tests, their Type I error and statistical power rates were similar through all the conditions of subgroups with equal number of studies. This suggests that when the studies are distributed equally within each category of the moderator the meta-analyst may apply any of the procedures in order to estimate the residual between-studies variance. Nevertheless, if the number of studies and the residual between-studies variances are roughly similar across categories, using a pooled estimate would be expected to provide more accurate results for most scenarios, as it takes into account a larger number of studies. This can be particularly important if the total number of studies is small (e.g., <20), which has been found to be the case for most Cochrane Reviews (Davey, Turner, Clarke, & Higgins, 2011).

When the number of studies was distributed unequally across categories, the practical consequences of having heterogeneous residual between-studies variances were more evident, with both tests leading to the wrong statistical conclusion more often than in the conditions with balanced subgroups. Specifically, under the condition of heterogeneity where the value of the smallest residual between-studies variance (τ^2 = 0.08) was associated with the category with the smallest number of studies, the $Q_{B(S)}$ test showed adequate control of the Type I error rate with at least 60 studies, whereas that the $Q_{B(P)}$ test yielded over-conservative Type I error rates and poor performance in terms of statistical power regardless of the number of studies. Under conditions where the value of the

Table 6. Average power rates by difference between the parametric mean effect size of each category of the moderator ($|\mu_{\delta 1} - \mu_{\delta 2}|$), by number of studies (*k*), by balanced and unbalanced distribution of *k*, and by the residual between-studies variance of each category of the moderator ($\tau_1^2 : \tau_2^2$)

	$Q_{B(S)}$	$Q_{B(P)}$
$\lvert\mu_{\delta 1} - \mu_{\delta 2}\rvert$		
0.2	0.2843	0.2783
0.4	0.7102	0.7095
K		
12	0.2674	0.2418
20	0.3359	0.3307
40	0.5179	0.5148
60	0.6378	0.6362
Distribution of *k*		
Balanced	0.5458	0.5464
Unbalanced	0.4729	0.4676
$\tau_1^2 : \tau_2^2$		
0.08: 0.08	0.5540	0.5530
0.16: 0.16	0.4453	0.4405
0.08: 0.16	0.5109	0.4711
0.16: 0.08	0.4787	0.5109

largest residual between-studies variance (τ^2 = 0.16) was associated with the category with the smallest number of studies, both tests provided inflated Type I error rates, with the $Q_{B(P)}$ test showing a greater departure from the nominal significance level. Note that the performance of the $Q_{B(P)}$ test was similar to that expected for the *F* test in a typical ANOVA with unbalanced sample sizes, when the homoscedasticity assumption was not met (Glass *et al.*, 1972; Hinkle *et al.*, 2003).

Lastly, our results also reflect that the $Q_{B(P)}$ test yielded more accurate control of error rates when the residual between-studies variances homogeneity assumption was fulfilled. In practice, the Q_B test is usually calculated using a pooled estimate (Borenstein *et al.*, 2009; Viechtbauer, 2010). Borenstein *et al.* (2009) and Viechtbauer (2010) suggested using a pooled estimate of the residual between-studies variance except when the meta-analyst suspects that the true value of the residual between-studies may vary from one category to the next.

As pointed out in the introduction, the most popular statistical packages for meta-analysis estimate the residual between-studies variance implementing only one of the two procedures described and compared throughout this paper, so that choice of software determines the method to be used. Our results showed some evidence that pooled or separate estimates might lead to a different performance of the Q_B test under some scenarios. Therefore, it would be helpful for the different meta-analysis software options to allow users to implement either method based on the characteristics of the database, as is already the case for Comprehensive Meta-Analysis 3.0 (Borenstein *et al.*, 2014). That would also allow undertaking sensitivity analyses if the meta-analyst suspects that the choice of procedure may have an impact on the results.

Results from our simulation study also shed some light on the accuracy of hypothesis testing for categorical moderators in meta-analysis, beyond the choice of pooled or separate variance estimates. The overall picture suggests that statistical tests can be expected to perform close to the nominal significance level in terms of Type I error, although greater between-studies variances and unbalanced category sizes may lead to inflated rates. Conversely, statistical power rates can be lower than desirable unless the difference among category effects and the number of studies are large enough. While the former may vary widely, the number of studies is often below 40 when the influence of a categorical moderator is statistically tested. Therefore, our results suggest that most of those analyses might be underpowered.

In conclusion, the results of our simulation study suggest that similar performance can be expected when using a pooled estimate or separate estimates of the residual between-studies variance to test the statistical association of a dichotomous moderator with the effect sizes, as long as there are at least 20 studies and these are roughly balanced across categories. Our results stress the need for a relatively large number of studies for the methods to have enough power to detect small to moderate differences among effect sizes from different subgroups. A pooled estimate will be preferable for most scenarios, unless the residual between-studies variances are clearly different and there are enough studies in each category to get precise separate estimates. Researchers are also encouraged to report the between-studies variance estimate(s) alongside its (their) confidence limits.

4.1. Limitations and future research

There are some limitations to this study. The results found can be generalized to the specific manipulated conditions. Although this study focused on standardized mean differences as the effect size index, our findings may be generalized to other effect size measures which follow an approximately normal distribution. In future simulation studies, it would be advisable to extend the manipulated conditions, for example, using other effect size indices, increasing the number of categories of the moderator and varying the average sample size of each meta-analysis.

In future studies, other estimators of the residual between-studies variance could be applied, such as the restricted maximum likelihood estimator (Viechtbauer, 2005), and they might also consider alternatives to the normal distribution to generate parametric effects, in order to mimic realistic scenarios more closely.

Finally, the Type I error and statistical power rates yielded by the methods considered in this study were suboptimal for many of the conditions examined. Previous simulation studies have demonstrated that the method proposed by Knapp and Hartung (2003) outperforms the standard method for testing the statistical significance of a continuous moderator (Viechtbauer, López-López, Sánchez-Meca, & Marín-Martínez, 2015). It would be interesting to evaluate the performance of this method to test for categorical moderators.

Acknowledgements

This article was supported by two grants from the Ministerio de Economía y Competitividad of the Spanish Government and by Fondo Europeo de Desarrollo Regional (Projects No. PSI2016-77676-P and PSI2015-71947-REDT).

References

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R., (2005). *Comprehensive meta-analysis* (Version 2.0) [Computer software]. Englewood, NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley. doi:10.1002/9780470743386

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2014). *Comprehensive meta-analysis* (Version 3.0) [Computer software]. Englewood, NJ: Biostat.

Borenstein, M., & Higgins, J. P. T. (2013). Meta-analysis and subgroups. *Prevention Science*, *14*, 134–143. doi:10.1007/s11121-013-0377-7

Davey, J., Turner, R. M., Clarke, M. J., & Higgins, J. P. T. (2011). Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: A cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, *11*, 160. doi:10.1186/1471-2288-11-160

DerSimonian, R., & Laird, N. (1986). Meta-analysis of clinical trials. *Controlled Clinical Trials*, *7*, 177–188. doi:10.1016/0197-2456(86)90046-2

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521–532. doi:10.1007/BF02293811

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston, MA: Allyn and Bacon.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed-effects analysis of variance and covariance. *Review of Educational Research*, *42*, 237–288.

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

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston: Houghton Mifflin.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*, 2693–2710. doi:10.1002/sim.1482

Konstantopoulos, S., & Hedges, L. V. (2009). Analyzing effect sizes: Fixed-effects models. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.) (pp. 279–293). New York, NY: Russell Sage Foundation.

López-López, J. A., Rubio-Aparicio, M., Sánchez-Meca, J., & Marín-Martínez, F. (2013, September). *Distribution of effect size and sample size in meta-analysis in the psychological field*. Paper presented at the XIII Congress of Methodology of the Social and Health Sciences, Tenerife, Spain.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.) (pp. 295–315). New York, NY: Russell Sage Foundation.

Review Manager (2014). *RevMan* (Version 5.3) [Computer software]. Copenhagen, Denmark: The Nordic Cochrane Centre, The Cochrane Collaboration.

Sánchez-Meca, J., Rosa-Alcázar, A. I., Marín-Martínez, F., & Gómez-Conesa, A. (2010). Psychological treatment of panic disorder with or without agoraphobia: A meta-analysis. *Clinical Psychology Review*, *30*, 37–50. doi:10.1016/j.cpr.2009.08.011

Senn, S. (2008). The t-test tool. *Significance*, *5*, 40–41. doi:10.1002/sim.3581

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*, 261–293. doi:10.3102/10769986030003261

Viechtbauer, W. (2010). Conducting meta-analysis in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. doi:10.18637/jss.v036.i03

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., & Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods*, *20*, 360–374. doi:10.1037/met0000023

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, *34*, 28–35. doi:10.1093/biomet/34.1-2.28

Received 15 May 2016; revised version received 2 December 2016

Appendix: Database for the example

Study	d	S_d	Random assignment
1	1.341	0.369	1
2	0.581	0.340	1
3	0.757	0.351	1
4	0.508	0.479	1
5	−0.023	0.558	1
6	0.044	0.277	1
7	0.428	0.270	1
8	0.819	0.521	1
9	−0.086	0.245	2
10	0.602	0.258	2
11	1.282	0.447	2
12	1.023	0.388	2
13	0.927	0.378	2
14	0.483	0.236	2
15	0.807	0.246	2
16	0.692	0.246	2
17	0.594	0.330	2
18	0.582	0.320	2
19	0.697	0.291	2
20	0.833	0.326	2
21	2.651	0.485	2
22	1.232	0.386	2
23	1.896	0.455	2
24	1.837	0.451	2
25	0.281	0.361	2
26	0.410	0.377	2
27	0.797	0.402	2
28	0.431	0.377	2
29	0.623	0.394	2
30	0.650	0.365	2
31	1.702	0.498	2
32	1.073	0.480	2
33	0.403	0.404	2
34	3.468	0.520	2
35	3.263	0.496	2
36	3.023	0.488	2
37	1.040	0.389	2
38	1.473	0.460	2
39	1.164	0.441	2
40	0.993	0.427	2
41	−0.344	0.381	2
42	−0.098	0.361	2
43	0.905	0.276	2
44	0.665	0.264	2
45	0.982	0.280	2

Continued

Appendix. (*Continued*)

Study	d	S_d	Random assignment
46	0.727	0.252	2
47	0.879	0.218	2
48	0.681	0.439	2
49	1.193	0.478	2
50	1.131	0.466	2

Notes. d = standardized mean difference for each study; S_d = standard error for the d index in each study. Random assignment = 1, no; 2, yes.
Source: Sánchez-Meca *et al.* (2010).