Journal of Educational and Behavioral Statistics
http://jebs.aera.net

Alternatives for Mixed-Effects Meta-Regression Models in the Reliability Generalization Approach: A Simulation Study

José Antonio López-López, Juan Botella, Julio Sánchez-Meca and Fulgencio Marín-Martínez

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2013 38: 443
originally published online 26 December 2012
DOI: 10.3102/1076998612466142

The online version of this article can be found at:
http://jeb.sagepub.com/content/38/5/443

Published on behalf of

American Educational Research Association

and

SAGE

http://www.sagepublications.com

Journal of Educational and Behavioral Statistics
2013, Vol. 38, No. 5, pp. 443–469
DOI: 10.3102/1076998612466142

Alternatives for Mixed-Effects Meta-Regression Models in the Reliability Generalization Approach: A Simulation Study

José Antonio López-López
University of Murcia, Spain

Juan Botella
Autonomous University of Madrid, Spain

Julio Sánchez-Meca
Fulgencio Marín-Martínez
University of Murcia, Spain

Since heterogeneity between reliability coefficients is usually found in reliability generalization studies, moderator analyses constitute a crucial step for that meta-analytic approach. In this study, different procedures for conducting mixed-effects meta-regression analyses were compared. Specifically, four transformation methods for the reliability coefficients, two estimators of the residual between-studies variance, and two methods for testing regression coefficients significance were combined in a Monte Carlo simulation study. The different methods were compared in terms of bias and mean square error (MSE) of the slope estimates, and Type I error and statistical power rates for the slope statistical tests. The results of the simulation study did not vary as a function of the residual variance estimator. All transformation methods provided negatively biased estimates, but both bias and MSE were reasonably small in all cases. In contrast, important differences were found regarding statistical tests, with the method proposed by Knapp and Hartung showing a better adjustment to the nominal significance level and higher power rates than the standard method.

Keywords: *reliability; reliability generalization; meta-analysis; coefficient alpha; mixed-effects meta-regression*

Reliability, as the consistency or reproducibility of the test scores (Crocker & Algina, 1986), is one of the most important psychometric properties to be considered when choosing a test for its administration in a specific context. However, reliability, such as it is defined and estimated from the classical test theory, is not a stable property for a given psychometric instrument, but rather a varying characteristic across

different applications of the test (Dawis, 1987; Gronlund & Linn, 1990; Pedhazur & Schmelkin, 1991). Thus, in order to obtain a reliability estimate representative enough for future test users, as well as to determine whether one or more factors from the sample characteristics or the administration context have an influence on the reliability of test scores, the best alternative is to quantitatively integrate the reliability coefficients computed with scores from different applications of the instrument under study. And, for carrying out a quantitative synthesis, meta-analysis constitutes an optimal methodological choice (Hedges & Olkin, 1985).

Despite previous meta-analytic studies integrating reliability coefficients that can be found in the literature (e.g., Churchill & Peter, 1984; Conway, Jako, & Goodman, 1995; Parker, 1983; Parker, Hanson, & Hunsley, 1988; Peter & Churchill, 1986; Salgado & Moscoso, 1996; Yarnold & Mueser, 1989), the term *reliability generalization* (RG) was first proposed by Vacha-Haase (1998). In an RG study, a set of reliability estimates from the same test are integrated, an overall reliability estimate is obtained, and heterogeneity between the individual reliability coefficients is assessed. Moreover, since some heterogeneity across estimates is usually found, a third objective in an RG study consists of looking for moderator variables in order to explain part of that variability.

Although Vacha-Haase's seminal paper was published just a few years ago, several dozen RG studies have already been carried out. A great variability can be found among these studies in terms of rigor, theoretical underpinning, and methodology. This is partially due to the fact that the RG approach was not conceived as monolithic in terms of the statistical methods applied (Henson & Thompson, 2002; Vacha-Haase, 1998; Vacha-Haase & Thompson, 2011). As a consequence, there is no consensus about several methodological issues affecting the statistical analyses.

One of these issues involves reliability coefficients transformation. Some authors advised to use untransformed reliability coefficients for the statistical analyses (e.g., Bonett, 2010; Henson & Thompson, 2002; Leach, Henson, Odom, & Cagle, 2006; Mason, Allam, & Brannick, 2007; Vacha-Haase, 1998). However, the distribution of the reliability coefficients is skewed for the most usual reliability measures (e.g., α coefficients and Pearson correlations), with a larger asymmetry level as the parameter approximates to one (Rodriguez & Maeda, 2006), as is usually the case for reliability coefficients reported in primary studies. Thus, some other authors recommended applying some transformation on the reliability coefficients in order to normalize their distribution and to stabilize their variances (e.g., Feldt & Charter, 2006; Rodriguez & Maeda, 2006; Sawilowsky, 2000). At least three different transformation formulas have been proposed and/or applied in the RG literature for coefficient α. These alternatives will be examined in more detail further below.

Another issue for which different solutions have been applied so far in the RG approach is the weighting scheme for the reliability coefficients. Some authors just employed ordinary least squares (OLS) analyses in their RG studies, that is, without weighting the reliability coefficients (e.g., Kieffer & Reese, 2002;

Leach et al., 2006, Vacha-Haase, 1998). Nonetheless, sample sizes in RG meta-analyses are usually unequal, leading to unequal sampling variances for the reliability coefficients, so that the homoscedasticity assumption—required for OLS techniques—is rarely met (Raudenbush, 1994; Rodriguez & Maeda, 2006). When weights were included in the analyses, some researchers chose the sample size as the weighting factor (e.g., Victorson, Barocas, & Song, 2008; Yin & Fan, 2000; Zangaro & Soeken, 2005), according to the proposal of Hunter and Schmidt (2004), while some others chose the inverse variance of the reliability coefficients (e.g., Aguayo, Vargas, de la Fuente, & Lozano, 2011; Beretvas, Suizzo, Durham, & Yarnell, 2008; López-Pina, Sánchez-Meca, & Rosa-Alcázar, 2009). Inverse variances have been used as the weighting factor in most of the meta-analyses published up to date (Borenstein, Hedges, Higgins, & Rothstein, 2010), and they are also becoming more and more frequent in the RG approach.

When the inverse variance is employed as the weighting scheme, it is necessary to assume some statistical model. One option is to assume a fixed-effect model. The fixed-effect model only considers one source of variation, the within-study variance, which refers to sampling error or the discrepancy between the reliability estimate and the value obtained if the instrument had been applied to the whole target population instead of the limited sample of subjects included in the study. An estimate of the within-study variance is required for the analyses, and several formulae are available for raw reliability coefficients as well as for the different transformations proposed in the literature. This implies that, once the transformation (or no transformation) of the reliability coefficients is chosen, then the estimation method for the sampling variance is unique. The fixed-effect model allows for generalizing results only to the samples whose reliability coefficients were included in the meta-analysis, and also to some external situations where the administration conditions and sample characteristics were identical (Hedges & Vevea, 1998).

An alternative is to assume a random-effects model, which is considered nowadays as the most realistic option in the general meta-analytic arena (Cooper, Hedges, & Valentine, 2009; National Research Council, 1992) and in the RG approach (Rodriguez & Maeda, 2006). The main reason for assuming a random-effects model is that, unlike the fixed-effect one, it allows for generalizing results beyond the studies included in the meta-analysis (Borenstein et al., 2010; Hedges & Vevea, 1998). While the fixed-effect model assumes a constant value for the parametric reliability coefficient across studies, the random-effects model assumes that the integrated reliability coefficients are estimating a random sample of parametric reliability coefficients extracted from a bigger superpopulation. In practice, that implies estimating a second variance component, the between-studies variance. Different procedures are available in the literature for estimating that variance (Viechtbauer, 2005), but no method will provide accurate results unless the number of studies integrated is large enough (Bonett, 2010; Borenstein et al., 2010). Due to the addition of a second variance component, the random-effects model provides more conservative results than the

fixed-effect model (Beretvas & Pastor, 2003). Nonetheless, several simulation studies have warned about the limitations of the fixed-effect model in general meta-analysis (e.g., Brockwell & Gordon, 2001; Marín-Martínez & Sánchez-Meca, 2010) and in the RG approach (Bonett, 2010; Romano & Kromrey, 2009).

The random-effects model has also some detractors. Bonett (2010) remarked that the sampling process of reliability coefficients in an RG meta-analysis is usually not random so that, strictly speaking, generalization to a population of reliability coefficients is not appropriate. Instead of the random-effects model, Bonett advocated the use of the varying-coefficient model, first proposed by Laird and Mosteller (1990), where parametric reliability coefficients are allowed to vary between studies but no random sampling is assumed, so that generalization is only intended to the set of primary studies included in the meta-analysis and some others with identical sample and administration conditions.

However, as stated by Laird and Mosteller (1990), "making inferences as if dealing with random samples contrary to fact is not a special issue for meta-analysis, but for all of science and technology" (p. 14). Although the majority of primary research in social, educational, and behavioral sciences is based on nonrandom samples, statistical inference techniques are routinely applied. The reason for this practice is that researchers usually consider their nonrandom samples as reasonably representative of the population to which inferences are intended to be made (Edgington, 1966; Frick, 1998; Overton, 1998). In the same vein, meta-analysts can make inferences to a larger population of studies as long as the set of selected studies can be considered a reasonably representative sample of that population, although the random sampling assumption is not strictly met (Schulze, 2004). Nonrandom sampling is, therefore, a characteristic of research and not specific of meta-analysis: "Rarely in research is the target population of samples fully enumerated and delimited; in fact, data sets used frequently consist of something close to convenience samples" (Schmidt, Oh, & Hayes, 2009, p. 103). The usual aim in meta-analysis, as in any kind of scientific research, is generalization of results and conclusions beyond the sample. In other words, "the goal of science is cumulative knowledge, and cumulative knowledge is generalizable knowledge" (Schmidt et al., 2009, p. 105). The random-effects model allows us for extending our conclusions beyond the set of studies included in the meta-analytic review (National Research Council, 1992). Therefore, the weighting scheme throughout this article will be that based on the inverse variance assuming a random-effects model.

Regarding moderator analyses, it is usually assumed that the predictors included in the model are fixed-effects variables. As a consequence, considering reliability coefficients as a random-effects variable leads to mixed-effects meta-regression models, where some study and sample characteristics are included as predictors of the variability between the reliability estimates. Moderator analyses constitute a crucial step in the RG approach (Rodriguez & Maeda, 2006), given the fact that most of the RG studies published so far found statistically significant relationships of one or more variables to the reliability

coefficients. As the psychometric theory predicts, several moderators associated to the variability of test scores have shown a statistically significant relationship with the reliability coefficients in many RG studies (e.g., standard deviation of test scores, type of population from which the sample subjects were recruited), and, for this reason, it has been argued that predictive models of the heterogeneity between reliability coefficients should always include some of them (Botella & Ponte, 2011). Other moderators that have proved a significant relationship with the reliability coefficients in previous RG studies are related to the test version (e.g., test length or original vs. adapted version).

When one or more predictors are included in the model, it becomes necessary to estimate the regression coefficients and, depending on the transformation applied to the reliability coefficients, these estimates will change to some extent. Also a new estimate of the (now residual) between-studies variance, which reflects the amount of heterogeneity on the dependent variable not explained by the moderators incorporated to the model, is required to be included into the weighting factor of mixed-effects analyses. Again, different procedures are available for computing that estimate, and the estimator of choice might have an influence on the results.

Apart from this, statistical tests for the regression model coefficients are required for testing the association of some moderator/moderators with the reliability estimates. The method traditionally computed in mixed-effects meta-analysis for addressing that issue has been criticized in the last few years, since its performance is strongly dependent on the accuracy of the variance estimates (Brockwell & Gordon, 2001). In order to solve that weakness, Knapp and Hartung (2003) proposed a new method based on the addition of a correction factor that takes into account the uncertainty of working with variance estimates instead of with known values. That proposal has not been employed yet in any published RG meta-analysis.

Purpose of This Study

Given the large number of methodological alternatives available when fitting mixed-effects meta-regression models in RG studies, the aim of the present article was to compare the performance of different combinations of methods under some realistic scenarios in RG studies by means of Monte Carlo simulation. Specifically, two estimators of the residual between-studies variance, two methods for testing the significance of the model regression coefficients, and four transformation methods (including untransformed reliability coefficients) were considered, leading to 16 methodological alternatives. Bias and efficiency were studied for the different estimation methods of the model regression coefficients, and Type I error and statistical power rates of their corresponding significance tests were then compared for all methodological combinations. Regarding outcome variables, coefficient α is the most widely reported reliability measure in primary studies and, since mixing different types of reliability

coefficients is not appropriate (c.f. Rodriguez & Maeda, 2006), most RG studies published so far have employed coefficient α as the main dependent variable. Thus, this study was focused on α coefficients (transformed or untransformed).

Regarding our hypotheses, we expected that the methods including the Knapp–Hartung correction would perform better than ones combined with the standard method, as reported by the authors in their seminal paper (Knapp & Hartung, 2003). Also, we expected that the transformed methods would outperform the untransformed ones, especially when comparing the Type I error and statistical power rates for the slope tests. Moreover, we expected small variations on the results depending on the residual between-studies variance estimator, as it was previously found (Knapp & Hartung, 2003).

The structure of this article runs as follows: First, the structure of a mixed-effects meta-regression model in the RG approach is sketched; second, the different methods compared in our study are presented; third, a review of previous simulation studies in the RG approach is summarized and then technical specifications of the current study are described; next, the study results are presented; and finally, the results are discussed and some conclusions are provided.

Mixed-Effects Meta-Regression Models in RG Studies

In an RG meta-analysis with k independent studies, let y denote a $(k \times 1)$ vector with k reliability coefficients $\{Y_i\}$, $\mathbf{X}$ a $[k \times (p + 1)]$ design matrix of full column rank (so that the inverse in Equation 2 exists) with p predictor variables. Then, a mixed-effects model is expressed with the formula (Raudenbush, 1994):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{u} + \mathbf{e}, \tag{1}$$

where $\mathbf{b}$ is a $[(p + 1) \times 1]$ vector containing the regression coefficients $\{\beta_0, \beta_1, \ldots, \beta_p\}$, $\mathbf{u}$ is a $(k \times 1)$ vector of independent between-studies errors $\{u_i\}$ with distribution $N(0, \tau^2)$, and $\mathbf{e}$ is a $(k \times 1)$ vector of independent within-study errors $\{e_i\}$, each of them with distribution $N(0, v_i)$. While v_i is the estimated within-study variance for the ith study, assumed as known in meta-analysis, τ^2 represents here the residual between-studies variance, that is, the remaining heterogeneity different to sampling error after adding one or more predictor variables to the model.

Regression coefficients $\{\beta_0, \beta_1, \ldots, \beta_p\}$ can be estimated using the weighted least squares formula:

$$\hat{\mathbf{b}} = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{W}}\mathbf{y}, \tag{2}$$

where $\hat{\mathbf{W}}$ is a $(k \times k)$ diagonal matrix with the inverse sampling variances as elements, that is, $\{1/(v_i + \hat{\tau}^2)\}$ under a mixed-effects model. Nearly unbiased estimators are available for v_i when using raw reliability coefficients as elements

TABLE 1.
Methods for Transforming Reliability Coefficients

Transformation	Coefficient	Back-transformation	Sampling Variance, v_i
Not transformed	$\hat{\alpha}_i = \frac{N_i}{N_i - 1}\left(1 - \frac{\mathbf{\Sigma}\sigma_q^2}{\sigma_X^2}\right)$	—	$V(\hat{\alpha}_i) = \frac{2J_i(1-\hat{\alpha}_i)^2}{(J_i-1)\left\{N_i - 2 - [(J_i-2)(k-1)]^{1/4}\right\}}$
Fisher's Z	$Z_i = \frac{1}{2}\log_e\left(\frac{1+\hat{\alpha}_i}{1-\hat{\alpha}_i}\right)$	$\hat{\alpha}_i = \frac{e^{2Z_i}-1}{e^{2Z_i}+1}$	$V(Z_i) = \frac{1}{N_i - 3}$
Hakstian and Whalen (1976)	$T_i = \sqrt[3]{1-\hat{\alpha}_i}$	$\hat{\alpha}_i = 1 - T_i^3$	$V(T_i) = \frac{18J_i(N_i-1)(1-\hat{\alpha}_i)^{2/3}}{(J_i-1)(9N_i-11)^2}$
Bonett (2002)	$L_i = \mathrm{Log}_e(1 - \lvert\hat{\alpha}_i\rvert)$	$\hat{\alpha}_i = 1 - e^{L_i}$	$V(L_i) = \frac{2J_i}{(J_i-1)(N_i-2)}$

Note: N_i = sample size for the ith sample; J_i = test length for the ith sample; σ_q^2 = variance of the scores in the qth item; σ_X^2 = variance of the total scores.

in **y**, and also when some transformation on the reliability coefficients is computed, as it will be detailed in the next section. Conversely, there are at least seven different estimators for τ^2 and none of them provides unbiased values. That issue will be discussed further below, since the estimator choice might have an influence on the results.

Methods for Transforming Coefficients α

According to the classical test theory (Crocker & Algina, 1986), reliability is defined as the quotient between the population variances of the true and observed scores, which can also be expressed as a squared correlation. Since true scores are unknown in practice, some alternative procedure for estimating the scores reliability is needed. Several methods are available nowadays for obtaining a reliability estimate. Since each of these methods is based on different theoretical assumptions, they will provide non-interchangeable reliability measures (Graham, 2006). Therefore, as they estimate different measurement errors, mixing different reliability coefficients in the same meta-analysis is not appropriate. Some of these methods (e.g., test–retest, parallel forms) compute the reliability coefficient as a correlation, so that it seems reasonable to transform these reliability coefficients using Fisher's Z, which was proposed for normalizing the distribution of Pearson correlations. Conversely, coefficient α is not a correlation, and it remains unclear which consequences should be expected after applying Fisher's Z transformation to these reliability coefficients.

Table 1 gathers some different options proposed in the RG literature for transforming coefficients α, including the raw α coefficients themselves on the first place. Up to date, when a transformation was applied, Fisher's Z was the most employed one, although that transformation is theoretically appropriate only when

the reliability coefficients are computed as a Pearson correlation, and that is not the case for α coefficients, which constitute the main dependent variable in almost every single RG study carried out so far. For that reason, Rodriguez and Maeda (2006) recommended using a transformation first proposed by Feldt (1969) for two samples and extended by Hakstian and Whalen (1976) for k samples. Finally, another transformation has been proposed more recently by Bonett (2002), in order to compensate for the fact that tests and confidence intervals for α are based on the usually unrealistic assumption that the k parts of the test are parallel. Bonett (2010) proposed to use this transformation on each individual coefficient α when fitting meta-regression models, as it is a variance-stabilizing and approximate normalizing transformation of the coefficient α distribution. In the same vein, Bonett's transformation can also be applied in the context of mixed-effects meta-regression models when α coefficients constitute the outcome variable. Table 1 provides formulae for computing all three transformations. Also, formulae for their respective sampling variances (v_i), under the normality assumption for the individuals' scores (Ray & Shadish, 1996), are presented. Note that the sampling variance presented in Table 1 for the Bonett's transformation was not proposed by him to be used as a weighting factor in meta-analysis. In fact, in the Bonett's (2010) varying coefficient model for meta-analysis, using weighting factors is unadvised. However, under a mixed-effects meta-regression model, using the Bonett's transformation and its sampling variance is a reasonable option.

When some transformation is applied, the statistical results are not directly comparable to those obtained using raw reliability coefficients as the dependent variable (Aguinis, Gottfredson, & Wright, 2011). To account for that issue, equations for back-transformation are gathered in Table 1. These formulae can be applied to mean coefficients α and their confidence limits, the intercept in a regression model, or the mean reliability values for each category in an analysis of variance (ANOVA).

In contrast, to back-transform regression slopes into the original metric, a different strategy is required, given that the value obtained by a simple back transformation of the slope could be misguiding when Y_i^T is not a linear transformation of coefficient α. Our proposal, based on the definition of the slope as the amount of change on the dependent variable as the predictor increases in one unit, is outlined below.

Let $Y_i^T = \beta_0^T + \beta_1^T X_i$ be a regression model where Y_i^T is a transformed reliability coefficient, β_0^T and β_1^T are the model coefficients expressed in the transformation metric, and X_i is a predictor. If X_i is set to values of 0 and 1, then two different predictions are obtained for the criterion, $Y_i^T[0]$ and $Y_i^T[1]$, and the slope is the difference between both predicted values, that is:

$$Y_i^T[1] - Y_i^T[0] = \beta_0^T + \beta_1^T(1) - \left[\beta_0^T + \beta_1^T(0)\right] = \beta_1^T. \quad (3)$$

In Equation 3, both the predicted values and the slope are in the metric of the transformation. Our proposal for reporting the slope in the metric of the original reliability coefficient, β_1^B, is to calculate the difference between the back transformations of the predicted values $Y_i^B[0]$ and $Y_i^B[1]$, using the corresponding formula from Table 1. (Note that this procedure provides a result different to the simple back-transformation of the slope.)

Residual Between-Studies Variance Estimators

There are at least seven different procedures for estimating the residual between-studies variance, as a result of extending estimators for the random-effects model (Sánchez-Meca & Marín-Martínez, 2008; Viechtbauer, 2005). In the present article, two estimators will be considered: the extension of the DerSimonian and Laird (DL) estimator and the extended restricted maximum likelihood (REML) estimator.

Extension of the DL estimator

The DL estimator is simple to obtain, since no iterative computation is required. Therefore, it has been the most frequently employed estimator for random-effects meta-analysis not only in the RG approach but also in any application field of meta-analysis. Moreover, it has been included in many different computerized tools developed to help researchers when integrating information from different studies. The extension of this estimator for a mixed-effects meta-regression model with k studies and p predictor variables (see Equation 1) is given by the formula (Raudenbush, 2009):

$$\hat{\tau}^2_{DL} = \frac{\mathbf{y}'\mathbf{P}\mathbf{y} - (k - p - 1)}{tr(\mathbf{P})}, \tag{4}$$

where tr denotes the trace of a matrix and $\mathbf{P}$ is obtained with the expression:

$$\mathbf{P} = \mathbf{W} - \mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}, \tag{5}$$

with elements $\{1/v_i\}$ for the diagonal matrix $\mathbf{W}$. When the resulting estimate is negative, it is truncated to zero.

Extension of the Restricted Maximum Likelihood Estimator (REML)

The REML estimator is a reasonable alternative to the DL procedure just presented. Although it requires iterative computation, this estimator has shown an appropriate performance in previous simulations under a random-effects model for a wide variety of conditions (Viechtbauer, 2005), and has also been recommended by Raudenbush (2009) for mixed-effects models. The estimating process consists of obtaining an adjustment value (a) and adding the previous estimate, $\hat{\tau}^2_0$,

so that a new estimate, $\hat{\tau}_1^2$, is then produced. That sequence is replicated until convergence between $\hat{\tau}_0^2$ and $\hat{\tau}_1^2$ is reached. The process can be expressed by:

$$\hat{\tau}_1^2 = \hat{\tau}_0^2 + a_{\text{REML}}. \tag{6}$$

Under a mixed-effects model, with p predictors included in the model (see Equation 1), the adjustment is given by (Raudenbush, 2009):

$$a_{\text{REML}} = \frac{\mathbf{y}'\hat{\mathbf{P}}\hat{\mathbf{P}}\mathbf{y} - tr(\hat{\mathbf{P}})}{tr(\hat{\mathbf{P}}\hat{\mathbf{P}})}, \tag{7}$$

where $\hat{\mathbf{P}}$ was defined in Equation 5 and $\{1/(v_i + \hat{\tau}^2)\}$ elements are employed in the diagonal matrix $\hat{\mathbf{W}}$. For the first iteration, $\hat{\tau}^2$ can be set to the value obtained with other estimator (e.g., the DL estimator). As in the DL estimator, negative values for the REML estimator are truncated to 0. Macros for computing both estimators have been developed for statistical packages such as R (Viechtbauer, 2010) and Stata (Harbord & Higgins, 2008).

Methods for Testing the Regression Coefficients Significance

The standard method for testing regression coefficients assumes a normal distribution for the regression model coefficient estimates. Despite its wide use in meta-analysis, some authors argued that this method does not take into account the uncertainty of working with estimated variances and that might produce misleading findings (Brockwell & Gordon, 2001; Van Houwelingen, Arends, & Stijnen, 2002). To offset that limitation, Knapp and Hartung (2003) developed a new method by incorporating a correction factor to the traditional formula. Also, their method assumes a *t*-distribution for the model coefficient values, instead of a normal distribution. Both procedures for testing regression coefficients significance will be presented in this section.

The Standard Method

According to the standard method, the variance–covariance matrix for regression coefficients is obtained with the expression:

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{STD}} = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}, \tag{8}$$

with elements $\{1/(v_i + \hat{\tau}^2)\}$ for the diagonal matrix $\hat{\mathbf{W}}$. The statistical test for each model coefficient, assuming a normal distribution, is then computed with:

$$Z_j = \frac{\hat{\beta}_j}{\sqrt{\hat{V}_{\text{STD}}(\hat{\beta}_j)}}, \tag{9}$$

with $\hat{\beta}_j$ being the $[j + 1]$ element of the $\hat{\mathbf{b}}$ vector, obtained with Equation 2, and $\hat{V}_{\mathrm{STD}}(\hat{\beta}_j)$ being the $[j + 1, j + 1]$ element of the $\hat{\mathbf{\Sigma}}_{\mathbf{STD}}$ matrix, computed with Equation 8. Although this has been almost the only method employed for testing coefficients from mixed-effects meta-regression models (not only in RG studies but also in other applications of meta-analysis), its adequacy is strongly dependent on the accuracy of the sampling variance estimates. Consequently, if those estimates were inaccurate, then the statistical conclusion provided by the standard method might not be correct (Knapp & Hartung, 2003; Sidik & Jonkman, 2005).

The Knapp–Hartung Method

Knapp and Hartung (2003) proposed the addition of a correction factor to Equation 8 in order to solve the problem mentioned above. Their proposal can be expressed with:

$$\hat{\mathbf{\Sigma}}_{\mathbf{KH}} = \mathbf{S}^2_{\hat{\mathbf{W}}}\left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}, \tag{10}$$

where $\mathbf{S}^2_{\hat{\mathbf{W}}}$ is computed from:

$$\mathbf{S}^2_{\hat{\mathbf{W}}} = \frac{(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})'\hat{\mathbf{W}}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})}{k - p - 1}, \tag{11}$$

$\hat{\mathbf{b}}$ is obtained with Equation 2, and $\{1/(v_i + \hat{\tau}^2)\}$ elements are included in the $\hat{\mathbf{W}}$ diagonal matrix. The statistical test, assuming a *t*-distribution for the regression coefficients, is then computed with the expression:

$$T_j = \frac{\hat{\beta}_j}{\sqrt{\hat{V}_{\mathrm{KH}}\left(\hat{\beta}_j\right)}}, \tag{12}$$

with $\hat{\beta}_j$ and $\hat{V}_{\mathrm{KH}}\left(\hat{\beta}_j\right)$ defined in Equations 2 and 10, respectively. Under the null hypothesis $\beta_j = 0$, it is assumed that T follows a *t*-distribution with $k - p - 1$ degrees of freedom, as the authors proposed. Note, however, that some other values for the degrees of freedom have been proposed (e.g., Berkey, Hoaglin, Mosteller, & Colditz, 1995).

In their simulation study, using log-odds ratios as the dependent variable, Knapp and Hartung (2003) found that their new method outperformed the standard one in terms of adjustment to the nominal significance level. Sidik and Jonkman (2005) obtained similar results when comparing both methods. However, the Knapp–Hartung proposal has not been implemented in any RG study yet and, since the dependent variable is different (reliability coefficients), its performance for RG meta-analyses is unknown. Macros for computing both methods have been developed for statistical packages such as R (Viechtbauer, 2010) and Stata (Harbord & Higgins, 2008).

TABLE 2.
Slope Estimates and Associated p *Values From the Example*

	DL			REML		
	$\hat{\beta}_1$	p_{STD}	p_{KH}	$\hat{\beta}_1$	p_{STD}	p_{KH}
Raw α coefficients	.0326	.064	.118	.0339	.079	.107
Fisher's *Z*	.0442	.257	.141	.0431	.140	.147
Hakstian–Whalen	.0424	.192	.125	.0412	.116	.112
Bonett	.0441	.284	.161	.0431	.155	.166

Note: $\hat{\beta}_1$ = slope estimate; DL, REML = extended DerSimonian and Laird and restricted maximum likelihood estimators for the residual between-studies variance; p_{STD}, p_{KH} = *p* values corresponding to the standard method and the Knapp–Hartung correction for testing the regression coefficients.

An Illustrative Example

An example is presented here in order to illustrate the 16 resulting methods after combining four alternatives for transforming reliability coefficients, two residual between-studies variance estimators and two methods for testing regression coefficients significance. Data for the example were extracted from an RG study about the Hamilton Rating Scale for Depression (the whole database is available in López-Pina et al., 2009). Considering the samples for which the 17-item version was administered, a meta-regression model was fit using each one of the proposed methods, with the score standard deviation, *SD,* as the predictor, and the coefficient α, Y_i, as the dependent variable. Table 2 gathers estimates for the model slope and *p* values for its statistical significance from each single analysis. When some transformation was applied on the reliability coefficients, slope estimates were back-transformed. Taking as a reference the combination of untransformed reliability coefficients with DL estimator, the regression equation was:

$$Y_i = 0.594 + 0.0326SD_i.$$

Regarding the results with the 16 procedures, the slope estimates were around .033 when the raw reliability coefficients were employed as the dependent variable, and values over .04 were obtained when using some transformation. The *p* values for the slope tests showed important discrepancies depending on the method considered. Assuming a 95% confidence level, statistically significant results were not achieved in any case; however, marginally significant results were found when applying the standard method for testing regression coefficients combined with raw α coefficients, for both DL and REML estimators (*p* values of .064 and .079, respectively). Conversely, the remaining methods provided *p* values greater than .10.

Therefore, as data from this example show, the method employed for fitting mixed-effects meta-regression models in the RG approach can affect the results, and that justifies a systematic comparison of the different methodological alternatives in order to determine which method is the most suitable one for a given scenario. In this study, we compared all of the aforementioned methods by means of Monte Carlo simulation. More details about that systematic comparison are provided below.

A Review of Previous Simulation Studies in the RG Approach

The RG approach constitutes a new application field for meta-analysis, and simulation studies are needed to assess the performance of the meta-analytic techniques in this framework, as well as to compare how the methodological alternatives specific to the RG field work under different scenarios. Regarding general meta-analytic methods, Mason et al. (2007) carried out a simulation study comparing the performance of different methods in mixed-effects models. However, the dependent variable in their study was the test–retest correlation instead of the α coefficient. Also, while these authors focused on the efficiency of the different methods included for estimating the model slope, in the present simulation bias, Type I error and statistical power rates were also considered as comparative criteria.

On the other hand, the fact that several transformations are available for the outcome variable is something specific of the RG approach. Feldt and Charter (2006) carried out a simulation study comparing different approaches for averaging internal consistency coefficients, some of them incorporating either the Fisher's *Z* or the Hakstian-Whalen transformations. Also, Bonett's transformation was employed in some recent simulation studies (Bonett, 2010; Romano & Kromrey, 2009). In the present study, all three transformations detailed along this paragraph for the reliability coefficients were included.

Simulation Study

A simulation was carried out to compare the 16 alternative methods for fitting mixed-effects meta-regression models presented above. The simulation was programmed in R, using *Metafor* (Viechtbauer, 2010) and *MCMCpack* (Martin, Quinn, & Park, 2011) packages. We decided to conduct this simulation under the classical test theory framework, because most of the tests chosen in previous RG studies were made based on the former.

Regarding manipulated factors in this simulation, sample sizes, N_i, were generated from a log-normal distribution with a mean value of 150 participants. The asymmetry of the sample size distribution was one of the conditions manipulated, with values of +1, +2, and +3, according to empirical asymmetry values observed in previous RG databases (e.g., Botella, Suero, & Gambara, 2010; López-Pina et al., 2009; Sánchez-Meca et al., 2011). Also, the number of studies

for each meta-analysis, k, was set to values of 15, 30, and 60. Finally, for the slope parametric value, two different scenarios were considered: For the first set of conditions, a predictor variable was generated from a distribution $N(0, 1)$ with no relationship to the reliability coefficients, so that the expected value for the slope was 0; for the second scenario, the error component in the test scores was generated as a function of that predictor, leading to a mean empirical slope, $\bar{\beta}_1$, of .01348 for all conditions (values between .01346 and .01350).

A key aspect in the simulation was the computation of the parametric coefficients α. In a first step, population test scores for each study were generated. Considering settings described in previous simulations (Bonett, 2010; Botella & Suero, 2012), a 20-item test was defined. For the calculation of each parametric coefficient α, a population of 10,000 subjects was defined. True scores for each of the 20 items, t_q, were generated from a multivariate normal distribution with mean 0, variance 2 for each item and covariance 0.4 for any pair of items. This provided a (10,000 $\times$ 20) matrix of true scores for each study. Then, error scores for each item, e_q, were generated from a normal distribution with mean 0, the variance changing from one study to the next due to the predictor value, with a range between .1 and 1.9. This resulted in another (10,000 $\times$ 20) matrix of error scores for each study. The observed scores for each of the 10,000 subjects in the qth item, x_{sq}, were calculated with the expression (Crocker & Algina, 1986):

$$x_{sq} = t_{sq} + e_{sq}.$$

Finally, scores for each subject in the whole test were computed as $x_s = \sum_{q=1}^{20} x_{sq}$. The parametric α coefficients, α_i, were computed from the database of 10,000 subjects generated for each study.

In a second step, samples of N_i subjects were taken from the respective populations, and the empirical α coefficients, the three proposed transformations, and their respective sampling variances were computed with the formulae gathered in Table 1. This process—generating a database of 10,000 subjects and then extracting a sample of N_i of them—was replicated k times in order to simulate the data corresponding to the k studies in an RG meta-analysis.

Once the sample reliability coefficients and within-study variances for the k studies were obtained, results for each meta-analysis were obtained by fitting mixed-effects meta-regression models for the 16 statistical alternatives under comparison. For each condition, 10,000 meta-analyses were computed.

Regarding comparative criteria, bias and mean square error (*MSE*) for the slope estimates were first computed in the conditions where $\beta_1 \neq 0$ for each one of the eight combinations (4 transformation methods $\times$ 2 residual variance estimators), providing different estimates of the model coefficients. When some transformation was applied on the reliability coefficients, the slope values were

TABLE 3.
Descriptive Statistics From the Simulated Data

Statistic	x_s	S_X^2	$\hat{\alpha}_i$
Minimum	−70.501	106.576	0.481
Maximum	65.646	328.786	0.818
Mean	0.003	212.211	0.719
Median	−0.007	210.336	0.720
Variance	211.966	796.081	0.001
Skewness	0.003	0.273	−0.989
Kurtosis	0.004	1.154	4.523

Note: x_s = observed total scores for each subject in a population of 1,000,000 participants; S_X^2 = variance of the total scores from each sample of participants; $\hat{\alpha}_i$ = empirical coefficient α for each sample of subjects.

back-transformed. Mathematical computations required for obtaining bias and *MSE* are provided below.

Let $\hat{\beta}_1^m$ be the slope estimate obtained with any of the proposed methods, and back-transformed to the reliability coefficients metric where necessary. The average of $\hat{\beta}_1^m$ for any given condition was computed with (Marín-Martínez & Sánchez-Meca, 2010):

$$\mathrm{AVE}\left(\hat{\beta}_1^m\right) = \frac{\sum \hat{\beta}_1^m}{10,000}. \tag{13}$$

Then, bias was obtained with:

$$\mathrm{BIAS}\left(\hat{\beta}_1^m\right) = \mathrm{AVE}\left(\hat{\beta}_1^m\right) - \bar{\beta}_1, \tag{14}$$

where $\bar{\beta}_1$ is the average of the empirical parametric slopes obtained along the 10,000 meta-analyses. On the other hand, *MSE* was calculated with:

$$MSE\left(\hat{\beta}_1^m\right) = \frac{\sum_i \left(\hat{\beta}_1^m - \bar{\beta}_1\right)^2}{10,000}. \tag{15}$$

Finally, the proportion of rejections of the null hypothesis $\beta_1 = 0$, assuming a 95% confidence level, was computed for all 16 combinations. That led us to compare the different methods in terms of Type I error rates for conditions where $\beta_1 = 0$, and in terms of statistical power when $\beta_1 \neq 0$.

Results

In order to illustrate the general trends of the simulated data, some descriptive statistics are presented in Table 3. Descriptives from the observed scores, x_s, were

obtained after generating a database of 1,000,000 scores. Next, data from 1,000 studies were simulated with an asymmetry index of 2 for the sample size distribution, computing for each study the score variance, S_X^2, and coefficient α estimate, $\hat{\alpha}_i$.

In the remainder of this section, the different methods described above will be assessed by means of the comparative criteria considered in the Monte Carlo simulation. First, accuracy of the slope estimates will be compared for the eight methodological alternatives (after combining four transformation methods and two residual variance estimators), in terms of bias and *MSE*. Then, performance of the slope statistical tests will be assessed for the 16 available alternatives (as a result of combining the 8 previous methods either with the standard method or with the Knapp–Hartung correction), in terms of Type I error and statistical power rates.

Accuracy of the Slope Estimates

Tables 4 and 5 present bias and *MSE* results, respectively, for the eight estimators of the meta-regression model slope. In order to facilitate their interpretation, values on both tables were multiplied by 10,000, so that the reference slope value is now 134.8.

Results in Table 4 show that all conditions provided negatively biased estimates of the slope parameter, although that bias was smaller than 3% for any combination of methods. Results were very similar and showed identical trends regardless of the residual variance estimator, but some differences were observed depending on the transformation method. Specifically, the Bonett transformation systematically showed the highest bias rates, with the largest percentage of bias, around -2.9%, when both the asymmetry in the sample size distribution and the number of studies were small. In contrast, raw α coefficients provided bias results slightly smaller than the methods involving some transformation of the reliability coefficients when the asymmetry was small, while the Fisher transformation led to the smallest bias for larger values in both the asymmetry and the number of studies.

Regarding efficiency, results in Table 5 show some interesting trends as well. *MSE*s were slightly higher for all the methods as the asymmetry values increased, but the number of studies had a bigger influence decreasing the *MSE*s for larger k values. Again, results were almost the same for both DL and REML estimators. Focusing on the transformation method, however, raw α coefficients provided the largest *MSE*s along all of the simulated conditions, while the smallest values were obtained when applying Bonett's transformation.

Performance of the Slope Statistical Tests

Table 6 gathers Type I error results, while statistical power rates are provided in Table 7. In both tables, only results for the DL estimator are presented, since

TABLE 4.
Bias in the Slope Estimates for the Different Combinations of Methods

k		15			30			60		
Asymmetry		1	2	3	1	2	3	1	2	3
Raw α coefficients	DL	−2.517	−0.587	−1.847	−1.962	−1.158	−0.669	−2.235	−1.999	−1.141
	REML	−2.552	−0.605	−1.802	−1.972	−1.138	−0.646	−2.238	−1.998	−1.138
Fisher's *Z*	DL	−2.904	−1.665	−2.412	−1.659	−0.852	−0.283	−1.401	−1.433	−0.668
	REML	−2.902	−1.611	−2.400	−1.660	−0.848	−0.275	−1.401	−1.434	−0.678
Hakstian–Whalen	DL	−3.184	−1.757	−2.767	−2.174	−1.412	−0.912	−2.129	−2.071	−1.305
	REML	−3.203	−1.719	−2.738	−2.183	−1.417	−0.895	−2.130	−2.071	−1.314
Bonett	DL	−3.850	−2.636	−3.460	−2.557	−1.773	−1.248	−2.297	−2.309	−1.555
	REML	−3.873	−2.534	−3.363	−2.565	−1.787	−1.281	−2.298	−2.318	−1.550

Note: *k* = number of studies; Asymmetry = skewness of the sample size distribution; DL and REML = extended DerSimonian and Laird and restricted maximum likelihood estimators for the residual between studies variance.

TABLE 5.
MSE in the Slope Estimates for the Different Combinations of Methods

k		15			30			60		
Asymmetry		1	2	3	1	2	3	1	2	3
Raw α coefficients	DL	.6545	.7341	.7651	.2900	.3050	.3209	.1355	.1413	.1438
	REML	.6545	.7341	.7647	.2899	.3049	.3207	.1355	.1413	.1439
Fisher's *Z*	DL	.6344	.6963	.7111	.2803	.2842	.2941	.1305	.1339	.1344
	REML	.6344	.6953	.7077	.2803	.2842	.2933	.1305	.1339	.1343
Hakstian–Whalen	DL	.6301	.6954	.7164	.2778	.2851	.2959	.1294	.1333	.1340
	REML	.6301	.6944	.7123	.2778	.2849	.2947	.1294	.1333	.1339
Bonett	DL	.6219	.6832	.7006	.2739	.2786	.2882	.1277	.1311	.1315
	REML	.6218	.6812	.6935	.2739	.2782	.2859	.1277	.1309	.1310

Note: *k* = number of studies; Asymmetry = skewness of the sample size distribution; DL and REML = extended DerSimonian and Laird and restricted maximum likelihood estimators for the residual between studies variance.

TABLE 6.
Type I Error Rates for the Slope Tests Using the DL Estimator

k		15			30			60		
Asymmetry		1	2	3	1	2	3	1	2	3
Raw α coefficients	STD	.017	.020	.021	.014	.018	.020	.017	.018	.021
	KH	.048	.053	.054	.049	.050	.050	.050	.053	.055
Fisher's *Z*	STD	.006	.007	.007	.006	.006	.006	.005	.005	.006
	KH	.046	.049	.049	.050	.048	.043	.048	.048	.048
Hakstian–Whalen	STD	.017	.020	.023	.017	.019	.019	.019	.020	.021
	KH	.046	.049	.049	.049	.049	.044	.047	.049	.048
Bonett	STD	.016	.020	.022	.018	.020	.019	.020	.020	.021
	KH	.046	.048	.049	.049	.047	.043	.048	.048	.047

Note: *k* = number of studies; Asymmetry = skewness of the sample size distribution; STD and KH = standard method and Knapp–Hartung correction for testing the regression coefficients.

rates obtained with the REML were very similar and totally comparable in terms of the observed trends.

Assuming a 95% confidence level, accurate results for each method should be around .05 when the slope parametric value is 0. Results presented in Table 6 show that the rejection rates for the standard method were clearly under the nominal significance level, with rates smaller than .01 for the Fisher transformation and around .02 for the remaining transformation procedures. In contrast, the

TABLE 7.
Statistical Power Rates for the Slope Tests Using the DL Estimator

k		15			30			60		
Asymmetry		1	2	3	1	2	3	1	2	3
Raw α coefficients	STD	.266	.271	.259	.561	.552	.556	.884	.871	.875
	KH	.369	.363	.347	.682	.666	.658	.942	.929	.925
Fisher's *Z*	STD	.171	.170	.166	.425	.418	.422	.810	.795	.799
	KH	.368	.364	.348	.684	.675	.667	.942	.931	.930
Hakstian–Whalen	STD	.283	.282	.271	.587	.577	.580	.902	.890	.892
	KH	.370	.363	.341	.685	.671	.661	.943	.932	.929
Bonett	STD	.284	.281	.270	.586	.578	.581	.901	.892	.894
	KH	.369	.362	.345	.684	.674	.664	.942	.931	.930

Note: k = number of studies; Asymmetry = skewness of the sample size distribution; STD and KH = standard method and Knapp–Hartung correction for testing the regression coefficients.

Knapp–Hartung correction performed close to the nominal level for all of the transformation methods and along all simulated conditions.

Regarding statistical power rates, Table 7 shows that the lowest rates were obtained when combining the standard method and the Fisher transformation. The Knapp–Hartung correction systematically led to higher power rates than those obtained for the standard method. Apart from that, all rates increased for larger k values, while the asymmetry showed a small inverse relationship to the results. Finally, power rates were slightly higher when the Knapp–Hartung correction was combined with some transformation of the α coefficients.

Discussion

The present study was focused on the analyses of continuous moderators by fitting mixed-effects meta-regression models using α coefficients as the outcome variable. Throughout this article, different equations for transforming reliability coefficients, as well as for computing their respective sampling variances and for back-transforming them to the original reliability coefficients metric, were presented (see Table 1). Extensions of the DL and REML estimators for the residual between-studies variance were also detailed (Equations 4 and 6), as well as the standard method for testing the regression coefficients and the adjustment proposed for Knapp and Hartung (2003) to the former (Equations 9 and 12). Performance for all the presented methods was compared by means of Monte Carlo simulation, where bias and *MSE* for the slope estimates, as well as Type I error and statistical power rates for the slope tests, were the comparative criteria considered.

Out of the different methodological issues implied in the several procedures compared in our article, the choice of the residual between-studies variance estimator (DL vs. REML) produced negligible differences in the trends, the changes observed in the results for the different conditions being very small. On the other hand, the transformation method of the reliability coefficients had some influence on the comparative criteria considered for this study. Finally, the method employed for testing the significance of regression coefficients (standard vs. Knapp–Hartung) showed a critical influence on the Type I error and statistical power results.

Regarding transformations, in terms of bias, all methods provided negatively biased estimates of the regression coefficients, although raw α coefficients showed results slightly better than the ones obtained when applying some transformation, especially when the asymmetry in the sample size distribution was small. Conversely, *MSE*s were higher for raw reliability coefficients than for any of the transformed methods. However, since bias results were always smaller than 3% regarding the slope parameter, and *MSE* values were also small and very similar from one method to another, the conclusion should be that all four transformation methods performed reasonably well in terms of bias and efficiency. Also, from a conceptual point of view, Fisher's *Z* transformation should not be used with coefficients α, as that transformation is only appropriate when the reliability coefficients were computed as a Pearson correlation coefficient (e.g., test–retest reliability). Therefore, for coefficients α, Hakstian and Whalen's (1976) and Bonett's (2002) transformations should be selected.

Considering now the two methods included here for testing the model coefficients, compared to the standard method, the Knapp–Hartung correction provided empirical Type I error rates closer to the nominal significance level, performing almost nominally for all combinations and under all of the simulated scenarios. Regarding statistical power, the Knapp–Hartung correction showed higher rates than the standard method regardless of the rest of conditions manipulated. These power rates were slightly higher when the Knapp–Hartung correction was combined with some transformation of the reliability coefficients. However, a noteworthy finding is that, when integrating 15 or 30 coefficients α, as was the case for some previous RG studies, power rates were considerably lower than the .80 boundary recommended by the scientific community. Thus, having a moderate to large number of reliability coefficients seems to be an important requirement when conducting moderator analyses in RG studies.

Usefulness and Limitations of the Findings Presented in This Article

The simulation study carried out in the presented article showed that, when fitting mixed-effects meta-regression models with one covariate, the slope estimates can be negatively biased, although usually that bias is not large enough to represent a threat for the results. Also, despite the fact that *MSE*s for these

estimates were smaller when some transformation on the reliability coefficients was applied, results were very similar when comparing different transformation methods, and *MSE*s decreased noticeably as the number of coefficients α increased. Thus, our results suggest that all transformation methods compared here perform similarly in terms of bias and efficiency of the model slope estimates, so that researchers conducting RG studies should pay more attention to some other criteria before making their decisions about the statistical methods implemented.

In contrast to the previous statement, significance tests for the slope did show important differences along the methodological alternatives compared here. According to our results, RG researchers should take into account that testing the model coefficients with the standard method may lead to a loss of statistical power, as Table 7 reflects, so that some moderators of the variability between reliability coefficients might not be identified in their RG studies unless they are integrating a large number of reliability coefficients. The Knapp–Hartung correction outperformed the standard method in terms of statistical power, with rates systematically greater than those obtained with the standard method, and showed Type I error rates closer to the nominal significance level.

Regarding limitations of the methods included in the present study, the fact that only mixed-effects models were considered here might be seen as problematic, since some other options are present in published RG studies. However, the purpose of the present article was not to assess the methodological choices implemented up to date, but rather to compare the best methodological alternatives for future studies, based on the main objectives in an RG study itself and on the current statistical alternatives to accomplish them. Since reliability is not a stable property for a given psychometric instrument (e.g., Crocker & Algina, 1986; Gronlund & Linn, 1990), the RG approach was proposed by Vacha-Haase (1998) as a way to integrate a set of reliability estimates from different applications of a test, and to guide expectations of potential test users about reliability with their sample characteristics and their administration context. That implies generalizing results to some other scenarios not necessarily identical to the ones accounted for in the RG study, and only random-effects models allow researchers for making such generalizations (cf. Beretvas & Pastor, 2003; Borenstein et al., 2010; Hedges & Vevea, 1998; Raudenbush, 2009; Sánchez-Meca, López-López, & López-Pina, in press; Schmidt et al., 2009).

Thus, random-effects models allow for generalizing results beyond the meta-analytic sample of studies and, for that reason, they are considered nowadays as the most suitable option for most meta-analyses (e.g., Cooper et al., 2009; Field, 2003, 2005; National Research Council, 1992; Schmidt, 2010; Schmidt et al., 2009). Note, however, that assuming a random-effects model is only justified if the meta-analyst can consider (on a reasonable basis) the set of studies retrieved for his or her review to be a representative sample of the population to which generalizations are intended. Assuming a random-effects

model when conducting moderator analyses leads to mixed-effects models, as the ones presented in this article. In addition to inverse variances, sample sizes can be considered as random-effects weights. However, results are not expected to be influenced by the choice of weights in a random-effects model, but rather by the transformation method in the reliability coefficients (Mason et al., 2007).

Also, as in any simulation, conditions manipulated in our study cannot account for the whole universe of scenarios present in RG studies already carried out or to be done in the future. As an illustration of that, some RG studies have integrated larger numbers of sampling reliability estimates than the ones considered here (e.g., Yin & Fan, 2000), and similar results to the ones presented here for 60 studies should be expected with smaller *MSE*s and an additional gain of statistical power. Moreover, generating sample size values from a log-normal distribution may be a reasonable approximation to the real situation in many RG meta-analyses (Mason et al., 2007), where most of the primary studies used small-to-moderate inpatient samples while a few ones applied the test as a screening instrument to large samples from general population. Increasing the asymmetry of the sample size distribution produced slightly higher *MSE*s and smaller statistical power rates, although that factor did not show a big influence on any of the criteria compared here.

Finally, the use of coefficient α in this simulation study, as well as in most of the RG studies published to date, leads to some noteworthy considerations. As Graham (2006) remarked, coefficient α is based on the essentially τ-equivalent measurement model. This implies that, when a coefficient α is computed, it is assumed that all items measure the same latent trait, although probably with a different degree of precision. Researchers estimating reliability with coefficient α, or retrieving α coefficients for carrying out an RG study, must be aware of this assumption, because its violation would directly affect the validity of the reliability estimates for a given test. The generating process of the item scores in our simulation, which was detailed above, fulfilled the requirements of the essentially τ-equivalent measurement model.

The RG approach was recently proposed (Vacha-Haase, 1998), with the aim of applying a methodology for quantitative synthesis, meta-analysis, to the purpose of obtaining a representative reliability value along different administrations of a given test, as well as identifying which factors can explain variability across the set of reliability estimates. The latter objective implies carrying out moderator analyses, and different alternatives for addressing that issue are available to the meta-analyst. Results of this study mainly suggest that, when a mixed-effects model is assumed for the moderator analyses in an RG study, the Knapp–Hartung correction for the statistical test of the model coefficients provides rates closer to the nominal significance level regarding Type I error, and higher power rates than the ones obtained for the standard method. Performance for that correction seems then promising in the RG approach, where it has not been applied to date.

Funding

The authors disclosed receipt of the following financial support for the research and/or authorship of this article: This research was supported by a grant for junior researchers from Fundación Séneca, Region of Murcia, Spain, and by the Ministerio de Ciencia e Innovación, Spanish Government, Project No. PSI2009-12172.

References

Aguayo, R., Vargas, C., de la Fuente, E. I., & Lozano, L. M. (2011). A meta-analytic reliability generalization study of the Maslach Burnout Inventory. *International Journal of Clinical and Health Psychology*, *11*, 343–361.

Aguinis, H., Gottfredson, R. K., & Wright, T. A. (2011). Best-practice recommendations for estimating interaction effects using meta-analysis. *Journal of Organizational Behavior*, *32*, 1033–1043.

Beretvas, S. N., & Pastor, D. A. (2003). Using mixed-effects models in reliability generalization studies. *Educational and Psychological Measurement*, *63*, 75–95.

Beretvas, S. N., Suizzo, M.A., Durham, J. A., & Yarnell, L. M. (2008). A reliability generalization study of scores on Rotter's and Nowicki-Strickland's Locus of Control Scales. *Educational and Psychological Measurement*, *68*, 97–119.

Berkey, C. S., Hoaglin, D. C., Mosteller, F., & Colditz, G. A. (1995). A random-effects meta-regression model for meta-analysis. *Statistics in Medicine*, *14*, 395–411.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*, 335–340.

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods*, *15*, 368–385.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*, 97–111.

Botella, J., & Ponte, G. (2011). Effects of the heterogeneity of the variances on reliability generalization: An example with the Beck Depression Inventory. *Psicothema*, *23*, 516–522.

Botella, J., & Suero, M. (2012). Managing heterogeneity of variances in studies of internal consistency generalization. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 8*, 71–80.

Botella, J., Suero, M., & Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods*, *15*, 386–397.

Brockwell, S. E., & Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, *20*, 825–840.

Churchill, G. A., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research*, *21*, 360–375.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology*, *80*, 565–579.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart & Winston.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology*, *34*, 481–489.

Edgington, E. S. (1966). Statistical inference and nonrandom samples. *Psychological Bulletin*, *66*, 485–487.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika*, *34*, 363–373.

Feldt, L. S., & Charter, R. A. (2006). Averaging internal consistency reliability coefficients. *Educational and Psychological Measurement*, *66*, 215–227.

Field, A. P. (2003). The problems in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics*, *2*, 105–124.

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods*, *10*, 444–467.

Frick, R. W. (1998). Interpreting statistical testing: Process and propensity, not population and random sampling. *Behavior Research Methods, Instruments, & Computers*, *30*, 527–535.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, *66*, 930–944.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching*, 6th ed. New York, NY: Macmillan.

Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika*, *41*, 219–231.

Harbord, R. M., & Higgins, J. P. T. (2008). Meta-regression in Stata. *The Stata Journal*, *8*, 493–519.

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

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.

Henson, R. K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "reliability generalization" studies. *Measurement and Evaluation in Counseling and Development*, *35*, 113–126.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting errors and bias in research findings* (2nd ed.). Newbury Park, CA: Sage.

Kieffer, K. M., & Reese, R. J. (2002). A reliability generalization study of the Geriatric Depression Scale. *Educational and Psychological Measurement*, *62*, 969–994.

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

Laird, N. M., & Mosteller, F. (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care*, *6*, 5–30.

Leach, L. F., Henson, R. K., Odom, L. R., & Cagle, L. S. (2006). A reliability generalization study of the Self-Description Questionnaire. *Educational and Psychological Measurement*, *66*, 285–304.

López-Pina, J. A., Sánchez-Meca, J., & Rosa-Alcázar, A. I. (2009). The Hamilton Rating Scale for Depression: A meta-analytic reliability generalization study. *International Journal of Clinical and Health Psychology*, *9*, 143–159.

Marín-Martínez, F., & Sánchez-Meca, J. (2010). Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement*, *70*, 56–73.

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*, *42*, 1–21.

Mason, C., Allam, R., & Brannick, M. T. (2007). How to meta-analyze coefficient-of-stability estimates. *Educational and Psychological Measurement*, *67*, 765–783.

National Research Council. (1992). *Combining information: Statistical issues and opportunities for research*. Washington, DC: National Academy Press.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods*, *3*, 354–379.

Parker, K. (1983). A meta-analysis of the reliability and validity of the Rorschach. *Journal of Personality Assessment*, *47*, 227–231.

Parker, K. C. H., Hanson, R. K., & Hunsley, J. (1988). MMPI, Rorschach, and WAIS: A meta-analytic comparison of reliability, stability, and validity. *Psychological Bulletin*, *103*, 367–373.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Peter, J. P., & Churchill, G. A. (1986). Relationships among research design choices and psychometric properties of rating scales: A meta-analysis. *Journal of Marketing Research*, *23*, 1–10.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York, NY: Russell Sage Foundation.

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.

Ray, J. W., & Shadish, W. R. (1996). How interchangeable are different estimators of effect size? *Journal of Consulting and Clinical Psychology*, *64*, 1316–1325.

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, *11*, 306–322.

Romano, J. L., & Kromrey, J. D. (2009). What are the consequences if the assumption of independent observations is violated in reliability generalization meta-analysis studies? *Educational and Psychological Measurement*, *69*, 404–428.

Salgado, J. F., & Moscoso, S. (1996). Meta-analysis of interrater reliability of job performance ratings in validity studies of personnel selection. *Perceptual and Motor Skills*, *83*, 1195–1201.

Sánchez-Meca, J., López-López, J. A., & López-Pina, J. A. (in press). Some recommended statistical analytic practices when reliability generalization studies are conducted. *British Journal of Mathematical and Statistical Psychology*. doi: 10.1111/j.2044-8317.2012.02057.x

Sánchez-Meca, J., López-Pina, J. A., López-López, J. A., Marín-Martínez, F., Rosa-Alcázar, A. I., & Gómez-Conesa, A. (2011). The Maudsley Obsessive-Compulsive Inventory: A reliability generalization meta-analysis. *International Journal of Clinical and Health Psychology*, *11*, 473–493.

Sánchez-Meca, J., & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods*, *13*, 31–48.

Sawilowsky, S. (2000). Psychometrics versus datametrics: Comment on Vacha-Haase's "reliability generalization" method and some EPM editorial policies. *Educational and Psychological Measurement*, *60*, 157–173.

Schmidt, F. L. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science*, *5*, 233–242.

Schmidt, F. L., Oh, I.-S., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*, 97–128.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Cambridge, MA: Hogrefe & Huber.

Sidik, K., & Jonkman, J. N. (2005). A note on variance estimation in random effects meta-regression. *Journal of Biopharmaceutical Statistics*, *15*, 823–838.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6–20.

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development*, *44*, 159–168.

Van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine*, *21*, 589–624.

Victorson, D., Barocas, J., & Song, J. (2008). Reliability across studies from the functional assessment of cancer therapy-general (FACT-G) and its subscales: A reliability generalization. *Quality of Life Research*, *17*, 1137–1146.

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.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48.

Yarnold, P. R., & Mueser, K. T. (1989). Meta-analyses of the reliability of Type A behaviour measures. *British Journal of Medical Psychology*, *62*, 43–52.

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement*, *60*, 201–223.

Zangaro, G. A., & Soeken, K. L. (2005). Meta-analysis of the reliability and validity of Part B of the Index of Work Satisfaction across studies. *Journal of Nursing Measurement*, *13*, 7–22.

Authors

JOSÉ ANTONIO LÓPEZ-LÓPEZ is a research assistant at the University of Murcia; Faculty of Psychology, Espinardo Campus, 30100-Murcia, Spain; e-mail: josealopezlopez@um.es; web site: www.um.es/metaanalysis. His research interests are in the methodology and applications of meta-analysis and statistical methods in general.

JUAN BOTELLA is a full professor at the Autonomous University of Madrid; Campus de Cantoblanco; Ivan Pavlov, 6; 28049 Madrid; Spain; e-mail: juan.botella@uam.es. His research interests are related to meta-analysis and psychological assessment.

JULIO SÁNCHEZ-MECA is a full professor at the University of Murcia; Faculty of Psychology, Espinardo Campus, 30100-Murcia, Spain; e-mail: jsmeca@um.es; web site: www.um.es/metaanalysis. His research interests are in the methodology of metaanalysis and statistical methods in general.

FULGENCIO MARÍN-MARTÍNEZ is an assistant professor at the University of Murcia; Faculty of Psychology, Espinardo Campus, 30100-Murcia, Spain; e-mail: fulmarin@um.es; web site: www.um.es/metaanalysis. His research interests are in the statistical models in meta-analysis and their application in the social, educational, medical, and behavioral sciences.

Manuscript received July 27, 2011
First revision December 2, 2011
Second revision April 1, 2012
Third revision May 29, 2012
Accepted July 31, 2012