British Journal of Mathematical and Statistical Psychology (2013), 66, 402–425

The British Psychological Society

www.wileyonlinelibrary.com

Some recommended statistical analytic practices when reliability generalization studies are conducted

Julio Sánchez-Meca*, José Antonio López-López and José Antonio López-Pina

University of Murcia, Spain

Precursors of the reliability generalization (RG) meta-analytic approach have not established a single preferred analytic method. By means of five real RG examples, we examine how using different statistical methods to integrate coefficients alpha can influence results in RG studies. Specifically, we compare thirteen different statistical models for averaging reliability coefficients and searching for moderator variables that differ in terms of: (a) whether to transform or not the coefficients alpha, and (b) the statistical model assumed, distinguishing between ordinary least squares methods, the fixed-effect (FE) model, the varying coefficient (VC) model, and several versions of the random-effects (RE) model. The results obtained with the different methods exhibited important discrepancies, especially regarding moderator analyses. The main criterion for the model choice should be the extent to which the meta-analyst intends to generalize the results. RE models are the most appropriate when the meta-analyst aims to generalize to a hypothetical population of past or future studies, while FE and VC models are the most appropriate when the interest focuses on generalizing the results to a population of studies identical to those included in the meta-analysis. Finally, some guidelines are proposed for selecting the statistical model when conducting an RG study.

1. Introduction

1.1. Background

Reliability is one of the most important properties when applying a test to a sample of participants. However, researchers do not usually report reliability estimates from their sample data. This is because researchers and clinicians tend incorrectly to consider reliability as a stable property of a measurement instrument. Nevertheless, since reliability is a property of the scores from one application of the test, it may vary from one sample to

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

DOI:10.1111/j.2044-8317.2012.02057.x

another if the composition, variability and administration conditions of the sample also vary (Crocker & Algina, 1986; Gronlund & Linn, 1990).

If reliability fluctuates for different applications of a test, then the best way to examine the measurement error variance for the test scores is to study its performance on a number of occasions. This requires a quantitative integration of reliability coefficients, and meta-analysis is the most suitable option for carrying this out. Vacha-Haase (1998) proposed the meta-analytic approach of *reliability generalization* (RG). An RG study is basically a meta-analysis where reliability coefficients obtained in different applications of a test are integrated, in order to compute an average reliability estimate and to find characteristics from studies and their samples which may be able to explain the heterogeneity of coefficients (Botella & Suero, 2012; Henson & Thompson, 2002; Rodriguez & Maeda, 2006; Vacha-Haase & Thompson, 2011). The RG approach is a kind of psychometric meta-analysis similar to the validity generalization approach proposed by Hunter and Schmidt (1990, 2004). Although these authors developed formulae to correct several statistical artefacts that affect validity coefficients (measurement error, range restriction, dichotomizing variables, etc.), RG studies do not usually apply such artefact corrections.

Nowadays, there is no single perspective concerning which statistical methods should be applied when conducting an RG study, giving freedom of choice to the meta-analysts (Botella & Ponte, 2011; Feldt & Charter, 2006; Henson & Thompson, 2002; Mason, Allam, & Brannick, 2007; Rodriguez & Maeda, 2006; Rouse, 2007; Vacha-Haase & Thompson, 2011). This freedom has not been considered as problematic by precursors of the RG approach. On the contrary, they encouraged researchers not to consider statistical analyses in RG studies in a single way. For example, in her seminal work, Vacha-Haase (1998) presented the RG method and applied it to a concrete example, concluding that 'different analytic tools than those illustrated here might be used in reliability generalization studies' (p. 16). Later, Thompson and Vacha-Haase (2000, p. 185) stated that 'we don't see RG as involving always a single genre of analyses'. Later, Henson and Thompson (2002, p. 124) established that 'because RG is not conceived as a monolithic method, there are a variety of ways in which an RG study could be conducted'. And Vacha-Haase, Henson, and Caruso (2002, p. 566) affirmed that 'RG is anything but monolithic regarding analytic choices', taking the RG studies existing to date to 'illustrate the potential for creativity in RG analysis and result presentation'.

The danger inherent in this *carte blanche* could arise if different conclusions were reached in an RG study depending on the selected statistical model. Some previous studies have dealt with this problem but, to our knowledge, no exhaustive comparisons have been carried out with the results obtained applying different methods and statistical models to the same set of studies. The developers of RG (e.g., Henson & Thompson, 2002; Vacha-Haase, 1998; Vacha-Haase *et al.*, 2002) were not advocating complete freewheeling with respect to how RG statistical analyses were conducted. Instead, they deemed it beneficial to explore various alternatives while the new RG methods initially evolved, but they realized that the field might eventually come to some consensus regarding preferred practices.

1.2. Purpose of the study

Our purpose was to assess the extent to which different statistical methods employed, or recently proposed, can lead to different results and conclusions when they are applied to the same data set of reliability coefficients. Specifically, our objectives were to compare the results obtained when applying different statistical methods in an RG study to average

a set of reliability coefficients and to test the influence of characteristics of the studies and samples on the heterogeneity exhibited by the reliability estimates. In doing so, we will present a conceptualization of the different statistical models for conducting an RG study. Next, different statistical methods will be applied to five real RG studies. Finally, some recommendations will be made regarding when the different statistical models should be assumed in a given RG study. It is important to note that, by comparing the results of the different statistical methods to a few real RG studies, it is not possible to determine which of the statistical models is the best. Our comparative study only intends to empirically examine whether applying different methods will lead to relevant discrepancies in the results of an RG study. Therefore, the recommendations made in Section 4 regarding which statistical model should be selected are based on an overview of the methodological work and simulation studies published in other sources. Finally, it is also important to note that this paper focuses on coefficient alpha, since it is the most commonly reported reliability coefficient in primary studies and, consequently, the main outcome measure in RG studies. Most of the guidelines and recommendations that are proposed here however, can also be applied when integrating other reliability coefficients (test–retest, parallel forms, etc.).

2. Statistical methods in RG studies

The main criterion for choosing a given statistical method in a meta-analysis is to assume, on a reasonable basis, the universe of potential studies to which the selected empirical studies pertain. Basically, three statistical models have been proposed to date in the meta-analytic arena (Borenstein, Hedges, Higgins, & Rothstein, 2010): the fixed-effect (FE) model (Hedges & Olkin, 1985; Konstantopoulos & Hedges, 2009), the random-effects (RE) model (Hedges & Vevea, 1998; Raudenbush, 2009), and the varying-coefficient (VC) model proposed by Laird and Mosteller (1990) and recently advocated by Bonett (2008, 2009, 2010). The consequences of assuming a given statistical model affect the extent to which the results can be generalized: FE and VC models aim to generalize the meta-analytic results only to studies with characteristics similar to those included in the meta-analysis, whereas the RE model is intended for generalizing to a broader superpopulation of studies. In addition, the different statistical methods proposed vary with regard to whether or not they transform and/or weight the reliability coefficients when they are statistically integrated. Next, we will present the main transformation methods used in RG studies to integrate coefficients alpha, as well as a conceptualization of the main statistical models applied or proposed in the RG literature.

2.1. To transform or not to transform

The first source of variability between the proposed RG methods is whether or not to apply a transformation of coefficients. Most RG studies to date have analysed untransformed coefficients alpha (e.g., Bachner & O'Rourke, 2007), as recommended by some authors (e.g., Henson & Thompson, 2002; Leach, Henson, Odom, & Cagle, 2006; Mason *et al.*, 2007; Thompson & Vacha-Haase, 2000). Other authors have encouraged meta-analysts to transform coefficients in order to normalize the distribution and/or stabilize the variances (e.g., Feldt & Charter, 2006; Rodriguez & Maeda, 2006; Sawilowsky, 2000).

When a transformation method has been applied in RG, the most common one has been the Fisher *z*-transform. Although this method is indicated only for reliability coefficients computed as a Pearson correlation (e.g., test–retest, parallel forms), the Fisher

z has been applied in RG studies mostly for transforming coefficients alpha (e.g., O'Rourke, 2004), or for transforming the square root of coefficient alpha (e.g., Graham & Christiansen, 2009). On other occasions, RG studies meta-analysed both untransformed coefficients alpha and their Fisher z-transform (e.g., Beretvas, Suizzo, Durham, & Yarnell, 2008).

For coefficient alpha, a more suitable transformation than the Fisher z is the one proposed by Hakstian and Whalen (1976) and recommended by Rodriguez and Maeda (2006). The Hakstian–Whalen transformation allows us to normalize the distribution of reliability coefficients. But, on a theoretical basis, a better transformation of coefficients alpha is the one proposed by Bonett (2002), as it enables us to normalize the distribution of coefficients alpha and to stabilize their variances. Although both the Hakstian–Whalen and Bonett transformations are theoretically more appropriate for coefficient alpha than the Fisher z, they have hardly been used in RG studies. On the other hand, although using the Fisher z to construct a confidence interval around coefficient alpha is theoretically inappropriate, a recent Monte Carlo simulation carried out by Romano, Kromrey, and Hibbard (2010) has shown that the Fisher z offers good performance and is very similar to the Bonett transformation in terms of empirical coverage probability. Some authors (e.g., Bonett, 2008, 2010; Hunter & Schmidt, 2004) have advised against using transformations for correlation coefficients and for coefficient alpha because these transformations produce biased estimates of the corresponding parameter, in particular when there is heterogeneity in the estimated coefficients. There is as yet, however, no evidence base upon which to rule out or advocate the use of transformations.

Table 1 presents the transformation formulae compared in our study. In particular, we included the raw coefficient alpha and the three aforementioned transformations. When some transformation is applied, Table 1 also shows formulae for back-transforming the average reliability estimate (and its confidence limits, if a confidence interval is computed) to the metric of alpha coefficients, in order to make the interpretation easier. Moreover, equations for computing the respective sampling variances, necessary for obtaining a confidence interval and carrying out meta-analytic methods when the inverse variance is used as the weighting factor, are provided.

2.2. Statistical models for RG studies

The various statistical models employed in RG studies differ not only on whether or not to transform the reliability coefficients, but also in the parameter(s) they are

Table 1. Transformation methods for coefficient alpha, with back-transformations and sampling variances

Transformation	Coefficient	Back-transformation	Sampling variance, $V(y_i)$
Not transformed	$\hat{\alpha}_i$	–	$V(\hat{\alpha}_i) = \frac{2J_i(1-\hat{\alpha}_i)^2}{(J_i-1)\{n_i-2-[(J-2)(k-1)]^{1/4}\}}$
Fisher z	$Z_i = \frac{1}{2}\ln\left(\frac{1+\hat{\alpha}_i}{1-\hat{\alpha}_i}\right)$	$\hat{\alpha}_i = \frac{e^{2Z_i}-1}{e^{2Z_i}+1}$	$V(Z_i) = \frac{1}{n_i-3}$
Hakstian–Whalen	$T_i = \sqrt[3]{1-\hat{\alpha}_i}$	$\hat{\alpha}_i = 1 - T_i^3$	$V(T_i) = \frac{18J_i(n_i-1)(1-\hat{\alpha}_i)^{2/3}}{(J_i-1)(9n_i-11)^2}$
Bonett	$L_i = \ln(1 - \lvert\hat{\alpha}_i\rvert)$	$\hat{\alpha}_i = 1 - e^{L_i}$	$V(L_i) = \frac{2J_i}{(J_i-1)(n_i-2)}$

Note. $\hat{\alpha}_i$: coefficient alpha for the ith study. n_i: sample size for the ith study. J_i: number of items of the test version used in the ith study. k: number of studies. The sampling variance for the untransformed coefficients alpha shown in the first row of the table is that proposed in Bonett (2010).

Table 2. Weighting schemes in the RG meta-analytic approach

Statistical method	Weighting scheme
Unweighted: OLS methods	$w_i = 1$
Inverse variance: FE model	$w_i^{FE} = \frac{1}{V(y_i)}$
Inverse variance: RE model	$w_i^{RE} = \frac{1}{V(y_i)+\hat{\tau}^2}$
Sample size, n_i: RE model	$w_i = n_i$

Note. y_i: reliability coefficient of the ith study (depending on the transformation method used, y_i can be the untransformed coefficients alpha, $\hat{\alpha}_i$, or the Z_i, T_i or L_i transformations). $V(y_i)$: sampling variance of y_i. n_i: sample size of the ith study. $\hat{\tau}^2$: between-studies variance estimate, defined in equation (1) (note that, for mixed-effects models, $\hat{\tau}^2_{Res}$, defined in equation (4), is employed instead of $\hat{\tau}^2$).

intended to estimate and in the weighting factor for analysing them. The four weighting methods usually applied in RG studies are: (1) unweighted, that is, applying ordinary least squares (OLS) techniques; (2) weighting by inverse variance, assuming an FE model; (3) weighting by inverse variance, assuming an RE model; and (4) weighting by sample size, following Hunter and Schmidt's (2004) validity generalization approach, which is also a kind of RE model. Table 2 presents formulae for the four weighting schemes.

Table 3 provides formulae for computing an overall reliability estimate and its sampling variance (required to construct confidence intervals around the average), with each of the statistical models to be presented in this section. When the coefficients alpha are very heterogeneous, then the weighted mean coefficient alpha is not representative of the set of reliability estimates and the next step in the analysis should be to search for characteristics of the studies that can explain at least part of that variability.

Methods for the computation of a statistical test and percentage of variance accounted for by a moderator variable on the reliability coefficients – using linear regression models – are also provided in Table 3.

2.2.1. Ordinary least squares methods

In her seminal work proposing the RG approach, Vacha-Haase (1998) applied conventional statistical methods that neither transformed nor weighted the reliability coefficients. Although this approach can be characterized as pertaining to the FE model (e.g., Mason *et al.*, 2007), it is presented separately. In Vacha-Haase's approach, an average coefficient alpha is obtained by calculating the simple arithmetic mean of the untransformed reliability estimates (Table 3). This method, also recommended by Henson and Thompson (2002) among others, is the most frequently used statistical model in RG studies (e.g., Bachner & O'Rourke, 2007).

The problem with this method is that the distribution of the coefficients alpha is highly skewed. Thus, some RG studies have previously translated coefficients alpha into Fisher's z (e.g., O'Rourke, 2004). This requires back-transforming to the metric of coefficient alpha (Table 1).

To search for study characteristics that can be associated with coefficients' variability, regression linear models can be applied (Table 3), taking the untransformed coefficient alpha or its Fisher z-transform as the dependent variable, and study characteristics as predictors.

Table 3. Statistical methods in the RG meta-analytic approach

Model	$\bar{y}$	$V(\bar{y})$	b	Σ_b	Test	R^2
OLS	$\bar{y}_u = \frac{\sum_i y_i}{k}$	$V(\bar{y}_u) = \frac{S_i^2}{k}$	$b_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$	$\Sigma_{b_{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}$	$t_j = \frac{b_j}{\sqrt{V(b_j)}}$	$R^2_{adj} = 1 - (1 - r^2_{xy})\frac{k-1}{k-2}$
FE	$\bar{y}_{FE} = \frac{\sum_i w_i^{FE} y_i}{\sum_i w_i^{FE}}$	$V(\bar{y}_{FE}) = \frac{1}{\sum_i w_i^{FE}}$	$b_{WLS} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$	$\Sigma_{b_{WLS}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$	$z_j = \frac{b_j}{\sqrt{V(b_j)}}$	$R^2 = \frac{\hat{\tau}^2 - \hat{\tau}^2_{Res}}{\hat{\tau}^2}$
VC	$\bar{\alpha}_u = \frac{\sum_i \hat{\alpha}_i}{k}$	$V(\bar{\alpha}_u) = \frac{\sum_i V(\hat{\alpha}_i)}{k^2}$	$b_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$	$\Sigma_{b_{VC}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$	$z_j = \frac{b_j}{\sqrt{V(b_j)}}$	$R^2_{adj} = 1 - (1 - r^2_{xy})\frac{k-1}{k-2}$
RE	$\bar{y}_{RE} = \frac{\sum_i w_i^{RE} y_i}{\sum_i w_i^{RE}}$	$V(\bar{y}_{RE}) = \frac{1}{\sum_i w_i^{RE}}$	$b_{WLS} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$	$\Sigma_{b_{WLS}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$	$z_j = \frac{b_j}{\sqrt{V(b_j)}}$	$R^2 = \frac{\hat{\tau}^2 - \hat{\tau}^2_{Res}}{\hat{\tau}^2}$
RE-c	$\bar{y}_{RE} = \frac{\sum_i w_i^{RE} y_i}{\sum_i w_i^{RE}}$	$V_{HA}(\bar{y}_{RE}) = \frac{\sum_i w_i^{RE}(y_i - \bar{y}_{RE})^2}{(k-1)\sum_i w_i^{RE}}$	$b_{WLS} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$	$\Sigma_{b_{KH}} = \frac{\mathbf{y}'\mathbf{P}\mathbf{y}}{k-p}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$	$t_j = \frac{b_j}{\sqrt{V(b_j)}}$	$R^2 = \frac{\hat{\tau}^2 - \hat{\tau}^2_{Res}}{\hat{\tau}^2}$
RE-n	$\bar{\alpha}_n = \frac{\sum_i n_i \hat{\alpha}_i}{\sum_i n_i}$	$V(\bar{\alpha}_n) = \frac{\sum_i n_i(\hat{\alpha}_i - \bar{\alpha}_n)^2}{k\sum_i n_i}$	$b_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$	$\Sigma_{b_{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}$	$t_j = \frac{b_j}{\sqrt{V(b_j)}}$	$R^2_{adj} = 1 - (1 - r^2_{xy})\frac{k-1}{k-2}$

Note. $\bar{y}$: average reliability estimate, either transformed or not ($\bar{\alpha}$). y_i: reliability coefficient of the ith study, either transformed or not ($\hat{\alpha}_i$). k: number of studies. $w_i^{FE}$and w_i^{RE}: FE and RE weights, both defined in Table 2. n_i: sample size of the ith study. $V(\bar{y})$: sampling variance of $\bar{y}$. $\mathbf{b}$: $(p \times 1)$-vector of regression coefficients with $p - 1$ predictors. $\mathbf{X}$: $(k \times p)$ design matrix. y: $(k \times 1)$-vector with elements $\{y_i\}$. $\mathbf{W}$: $(k \times k)$ diagonal matrix containing the weights (see Table 2). $\Sigma_{\mathbf{b}}$: $(p \times p)$ variance–covariance matrix of the regression coefficients. $\mathbf{V}$: $(k \times k)$ diagonal matrix with the sampling variances (for the Bonett-transformed coefficients). Test: statistical test for the moderator. b_j: jth regression coefficient from $\mathbf{b}$. $V(b_j)$: (j,j)th element from $\Sigma_{\mathbf{b}}$. R^2: proprtion of variance in the reliability coefficients accounted for by the moderator. r_{xy}: coefficient of correlation between the reliability estimates and the moderator. $\mathbf{P}$: matrix defined in equation (5), but now with $\{w_i^{RE}\}$ elements in $\mathbf{W}$. $\hat{\tau}^2$ and $\hat{\tau}^2_{Res}$: total and residual between-studies variance estimates, defined in equations (1) and (4), respectively. S_i^2: unbiased variance of y_i. [Note: Correction added on 21 June 2013 after initial publication online on 10 October 2012. The variance of the mean coefficient under OLS (third column, first row in the table) mistakenly appeared as $V(\bar{y}_u) = S_i^2$, and it should read $V(\bar{y}_u) = \frac{S_i^2}{k}$; this error has been corrected in this version of the article. All calculations presented in the article were carried out with the correct formula. Therefore, no other corrections were needed in the article.]

2.2.2. The fixed-effect model

The FE model defined on a set of k independent coefficients alpha assumes that all of them are estimating a common population coefficient alpha, α, that is, $\alpha_1 = \alpha_2 = \ldots = \alpha_i \ldots = \alpha_k = \alpha$. The only source of uncertainty assumed in the estimates, $\hat{\alpha}_i$, is the sampling error due to the fact that each sample is composed of different individuals. The underlying statistical model is given by $\hat{\alpha}_i = \alpha + u_i$, where u_i is the sampling error of $\hat{\alpha}_i$. The FE model, also called the common-effect model (Borenstein *et al.*, 2010), assumes that the samples of participants of the studies integrated are identical in composition and variability, and that the purpose of the meta-analyst is to generalize the results to a population of studies with identical characteristics to those included in the RG study.

Based on Hedges and Olkin (1985), (see also Hedges, 1994; Konstantopoulos & Hedges, 2009), it is possible to apply an FE model to statistically integrate a set of reliability coefficients. This proposal requires a previous transformation of the coefficients. Rodriguez and Maeda (2006) presented this approach by applying the Hakstian–Whalen transformation method. In the same vein, Bonett's transformation could also be applied from the FE model and, although theoretically inappropriate, the Fisher z-transform has been used as well when integrating coefficients alpha under an FE model (e.g., Zangaro & Soeken, 2005). Table 1 presents the transformation formulae, while the weighting scheme and statistical methods for the FE model are provided in Tables 2 and 3, respectively. Since some transformation is applied on the coefficients, the FE model requires back-transformation to the metric of coefficient alpha (Table 1).

If the meta-analyst considers that a few moderator variables can explain reliability coefficients variability, then statistical methods based on the FE model can be applied. Although some specific procedures are also available for categorical moderators (Hedges & Olkin, 1985), the influence of both categorical and continuous moderators on the reliability coefficients can be examined by applying linear regression models, assuming weighted least squares (WLS) as shown in Table 3.

2.2.3. The varying-coefficient model

Although several RG studies have applied statistical methods based on the FE model presented above, the assumption that all studies are estimating the same population coefficient alpha, α, is not very realistic. It is more reasonable to think that, as reliability of test scores changes according to the composition and variability of the samples, coefficients alpha from different studies are estimating different population reliability coefficients. The VC model was firstly proposed in meta-analysis by Laird and Mosteller (1990) and more recently advocated in RG by Bonett (2010). The VC model is a kind of FE model where it is assumed that the coefficient obtained in each study is estimating its own population reliability coefficient, that is, $\hat{\alpha}_i = \alpha_i + u_i$. Both models coincide in that the results are generalized only to studies identical to those included in the meta-analysis.

Statistical methods for the VC model are listed in Table 3. A simple arithmetic mean is computed for the overall reliability estimate, since the average of the k population reliability coefficients, $\bar{\alpha} = k^{-1} \sum_i \alpha_i$, is the parameter to be estimated. Although some specific methods have been proposed for categorical moderators (Bonett, 2010), linear regression models can be employed for analysing the influence of both categorical and continuous moderators, taking as dependent variable Bonett's transformation of coefficients alpha (Table 1).

2.2.4. The random-effects model

When the meta-analyst considers that the coefficients alpha obtained in a set of k studies constitute a reasonably representative sample of a hypothetical population of potential studies, then an RE model can be assumed to estimate the parameters of such a superpopulation of studies (Beretvas & Pastor, 2003; Hedges & Olkin, 1985; Hedges & Vevea, 1998; Overton, 1998; Rodriguez & Maeda, 2006). The RE model assumes that there is a hypothetical population of parametric coefficients alpha with mean μ_α and variance τ^2, focusing on obtaining estimates of μ_α and τ^2. The parametric coefficients alpha are defined as $\alpha_i = \mu_\alpha + e_i$, with e_i being the deviations of each population coefficient alpha, α_i, with respect to its population mean, μ_α, and are assumed to be normally distributed, $e_i \sim N(0, \tau^2)$. Theoretically, the RE model assumes that the k studies in an RG study have been randomly selected from a clearly defined superpopulation of potential studies, and then random samples of individuals are selected in each study to calculate a reliability coefficient. Each coefficient is, therefore, estimating its own population parameter, that is, $\hat{\alpha}_i = \alpha_i + u_i$. Consequently, RE models assume two sources of variability in the reliability estimates ($\hat{\alpha}_i = \mu_\alpha + e_i + u_i$): the within-study variability, due to the sampling of the participants in the samples, and the between-studies variability, due to assuming the existence of a superpopulation of parametric reliability coefficients from which the k studies constitute a representative sample of coefficients alpha.

As coefficients alpha have a non-normal distribution, RE model proponents recommend applying some transformation, such as the Fisher z-transform (Campbell, Pulos, Hogan, & Murry, 2005), Hakstian–Whalen transformation (Aguayo, Vargas, de la Fuente, & Lozano, 2011), or Bonett's transformation (Rendina-Gobioff, 2004).

In RE models, weights take into account both within-study $V(y_i)$ and between-studies variances, $\hat{\tau}^2$. As the between-studies variance is unknown, it has to be estimated from the studies that report reliability estimates. There are several methods available for estimating τ^2 (Raudenbush, 2009; Sánchez-Meca & Marín-Martínez, 2008; Viechtbauer, 2005). One of the most frequently used is an estimator based on the method of moments (DerSimonian & Laird, 1986), by means of

$$\hat{\tau}^2 = \frac{Q - (k-1)}{c}, \tag{1}$$

k being the number of studies, c given by

$$c = \sum_i w_i^{FE} - \frac{\sum_i (w_i^{FE})^2}{\sum_i w_i^{FE}}, \tag{2}$$

with w_i^{FE} defined in Table 2, and Q being a heterogeneity statistic defined by

$$Q = \sum_i w_i^{FE} (y_i - \bar{y}_{FE})^2, \tag{3}$$

with $\bar{y}_{FE}$ defined in Table 3. The Q statistic is frequently used to assess heterogeneity in meta-analysis (e.g., Hedges & Olkin, 1985).

When $\hat{\tau}^2$ is negative, it is truncated to zero. Table 3 provides equations for computing an average reliability coefficient and its sampling variance when assuming RE models.

Since some transformation is applied on the coefficients, back-transformation formulae are required (Table 1).

Both RE and VC models assume that coefficients alpha calculated in the studies are estimating their own population coefficient alpha, that is, both models assume that the population parameters are heterogeneous. The difference between the two models is that RE models assume that each study has been randomly selected from a superpopulation of potential studies and, as a consequence, it is possible to estimate the mean coefficient alpha of the superpopulation of studies, μ_α. Conversely, the VC model contends that, if the studies have not actually been randomly selected from a superpopulation of studies, then the coefficients alpha included in the meta-analysis are only representing their own population reliability coefficients, and the average of these population reliabilities, $\bar{\alpha} = \sum_i \alpha_i / k$, is the only parameter to be estimated.

Mixed-effects models can be applied for moderator analyses (taking study characteristics as fixed-effects variables). Statistical tests are available for categorical moderators (Hedges & Olkin, 1985), but the influence of both categorical and continuous moderators can be assessed by applying linear mixed-effects regression models (Table 3). Mixed-effects models required modifying weights, defined now as the inverse of the sum of the within-study variance, $V(y_i)$, shown in Table 1, and an estimate of the residual between-studies variance, $\hat{\tau}^2_{Res}$. One of the various different methods for estimating this parameter is based on the method of moments (DerSimonian & Kacker, 2007):

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

$\mathbf{y}$ being a vector of $k \times 1$ transformed coefficients (Table 1), tr($\mathbf{P}$) being the trace of $\mathbf{P}$, and $\mathbf{P}$ defined as (Raudenbush, 2009, p. 311):

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

where $\mathbf{W}$ is a diagonal $k \times k$ matrix with elements $\left\{w_i^{FE}\right\}$, defined in Table 2. When $\hat{\tau}^2_{Res}$ is negative, it is truncated to zero.

2.2.5. Corrected random-effects model

Conventional RE methods assume that sampling variances are known, although they have to be estimated in practice. If these estimates, included in the weights, were not accurate, then meta-analytic results and conclusions might not be correct. In order to take into account the uncertainty of working with estimates instead of with known values, Hartung (1999) proposed assuming a Student t distribution, instead of the standard normal distribution assumed for the conventional RE model, and using another formula for the sampling variance of the overall reliability estimate. His proposal outperformed the conventional method in a simulation study carried out by Sánchez-Meca and Marín-Martínez (2008). Similarly, Knapp and Hartung (2003) proposed a t-test for moderator analyses, and a correction factor for the variance–covariance matrix of the regression coefficients. Their refinement also showed good performance when compared to the standard method by simulation (Sidik & Jonkman, 2005). Formulae for the corrected RE model (RE-c) are presented in Table 3.

2.2.6. Weighting by sample size

This model, explicitly recommended by Yin and Fan (2000) for carrying out RG studies, implies analysing untransformed coefficients alpha and weighting them by sample size, in order to give more weight to the studies with larger sample sizes. Formulae for computing an overall reliability estimate and its sampling variance for this method, RE-n, are given in Table 3. This method is based on the Hunter–Schmidt meta-analytic approach and constitutes a kind of RE model (Hunter & Schmidt, 2004).

Table 3 presents OLS methods for this model when computing moderator analyses, as recommended by Hunter and Schmidt (2004) and implemented in some RG studies that weighted by sample size to calculate an average reliability estimate (e.g., Victorson, Barocas, & Song, 2008). Some other studies also weighted by sample size when conducting moderator analyses (e.g., Yin & Fan, 2000).

3. Some illustrative examples

If different statistical models applied to the same meta-analytic database lead to different conclusions, then selecting the statistical model is an important decision when conducting RG studies. In order to examine how different statistical models can lead to discrepant results, we applied the alternative statistical models presented above to five real RG studies. Note that with these comparisons we do not intend to find the best statistical model, but simply to illustrate the extent to which different methods can lead to important discrepancies in their results. The five RG studies compared here focused on the following psychological tests: the Maudsley Obsessive-Compulsive Inventory, MOCI (Sánchez-Meca, López-Pina, López-López, Marín-Martínez, Rosa-Alcázar, & Gómez-Conesa, 2011), State-Trait Anxiety Inventory, STAI (Botella, Suero, & Gambara, 2010), Yale-Brown Obsessive-Compulsive Scale, Y-BOCS (López-Pina *et al.*, 2010; July), Hamilton Rating Scale for Depression, HAM-D (López-Pina, Sánchez-Meca, & Rosa-Alcázar, 2009), and Maslach Burnout Inventory, MBI (Aguayo *et al.*, 2011).

A total of 13 different statistical models were compared: (1) OLS methods applied both on untransformed coefficients alpha and on Fisher's z-transform; (2) FE model applied on Fisher's z, Hakstian–Whalen, and Bonett transformations; (3) VC model; (4) RE model applied on Fisher's z, Hakstian–Whalen, and Bonett transformations; (5) RE-c model applied to the aforementioned three transformations; and (6) RE-n model. These models were selected because they have been used in some RG studies or have been recently proposed in the meta-analytic literature (although they may not yet have been applied). Statistical analyses were programmed in R, using the *metafor* package (Viechtbauer, 2010).

Note that combining the four transformation methods of coefficient alpha (untransformed, Fisher's z, Hakstian–Whalen, and Bonett; see Table 2) and six statistical models (OLS, FE, VC, RE, RE-c, and RE-n; see Table 3), a total of 24 statistical methods could be proposed. However, several of these combinations are theoretically inadequate. For instance, the VC model proposed by Bonett in the RG approach cannot be combined with the four transformation methods, as Bonett's proposal was very specific: not to transform the coefficients when calculating an average coefficient alpha and to apply Bonett's transformation when adjusting linear models to search for moderator variables. The RE-n model, on the other hand, is an extension of Hunter and Schmidt's meta-analytic approach, and in this model it is not recommended to transform the coefficients. In addition, the FE, RE, and RE-c models have been proposed in the literature together with some transformation of coefficients alpha to normalize their distribution. As a

consequence, these three models were combined with the three transformation methods presented here. Finally, the OLS model has been applied in the RG arena with the untransformed coefficients alpha, transforming them into Fisher's z. Thus, although the OLS model could also be applied with Hakstian and Whalen's and Bonett's transformations, we only included in our comparison study the two options that have actually been applied in previous RG studies.

With each database we calculated an average coefficient alpha and a confidence interval from each statistical model. As a comparison criterion, we defined a discrepancy index,

$$D = \frac{\bar{y}_j - \bar{y}_c}{\bar{y}_c} \times 100, \tag{6}$$

$\bar{y}_c$ being the reference mean. Since it has been the method most usually applied in RG studies, we selected the unweighted average of untransformed coefficients as reference. On the other hand, $\bar{y}_j$ represents each of the 12 remaining average estimates. Estimates obtained using some transformation method were back-transformed to coefficient alpha by the corresponding inverse formulae to make the values comparable (Table 1). Note that the simple average of untransformed and unweighted coefficients alpha, $\bar{\alpha}_u$, is the estimator proposed for both OLS and VC methods. On a reasonable basis, we considered discrepancies greater than or equal to 5% as reflecting practically significant differences among the procedures. OLS, FE, and VC methods all estimate the average of the population coefficients alpha included in the meta-analysis ($\bar{\alpha}$), whereas RE models estimate a different parameter, that is, the average superpopulation coefficient alpha, μ_α.

Confidence intervals were computed, and then compared in terms of the confidence width once back-transformed to the metric of coefficient alpha. The general formula for computing them was

$$\bar{y} \pm |z_{\alpha/2}| \sqrt{V(\bar{y})}, \tag{7}$$

with $\bar{y}$ and $V(\bar{y})$ obtained for each method with the respective formulae in Table 3, and $z_{\alpha/2}$ being the 100(α/2)% standard normal distribution score. For OLS and RE-c methods, the 100(α/2)% score in a t-distribution with $k - 1$ degrees of freedom, $_{\alpha/2}t_{k\text{-}1}$, was employed instead of $z_{\alpha/2}$.

For the VC model, a more specific formula has been proposed (Bonett, 2010):

$$1 - \exp\left[\ln(1 - \bar{\alpha}_u) - b \pm |z_{\alpha/2}| \sqrt{V(\bar{\alpha}_u)/(1 - \bar{\alpha}_u)^2}\right], \tag{8}$$

where $b = \ln\left[\bar{n}/(\bar{n} - 1)\right]$ is a correction factor for the slight positive bias of $\ln(1 - \bar{\alpha})$, and $\bar{n} = k/\sum_i (1/n_i)$ is the harmonic mean of the sample sizes of the studies, while $\bar{\alpha}_u$ and $V(\bar{\alpha}_u)$ were defined in Table 3.

From the statistical theory, confidence intervals around the mean coefficient alpha are expected to be narrower with FE models, followed by VC and, lastly, RE and OLS models.

Finally, the influence of moderators was compared with the p-values and proportions of variance accounted for in the statistical tests from the alternative models. For the latter, the R^2 index was used, reporting the adjusted R^2 index for OLS, VC, and RE-n methods. For

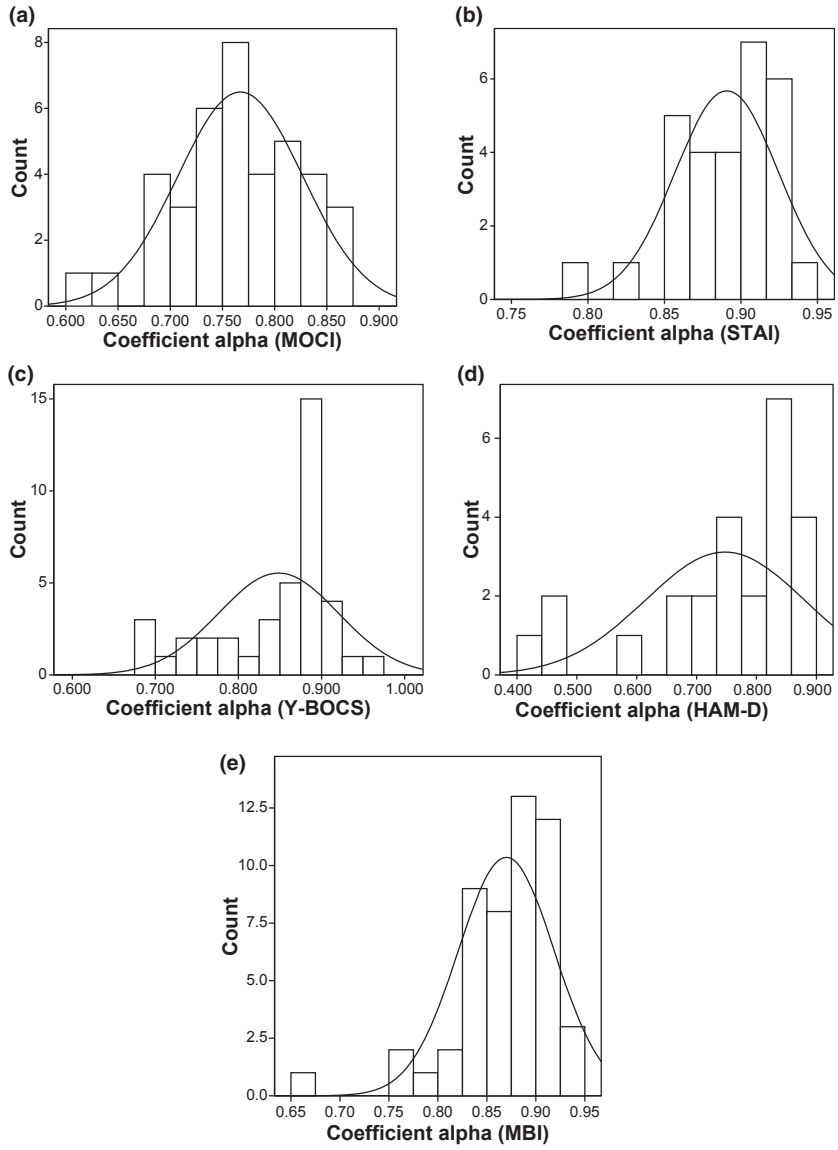

Figure 1. Histogram of distribution of coefficients alpha for the five RG examples.

the remaining methods, the proportion of between-studies variance removed after including predictors in the model was computed, as recommended by Raudenbush (1994, 2009; see also Aloe, Becker, & Pigott, 2010) for RE models and, although the between-studies variance is not a reference parameter in the FE model, the same rationale was employed for computing an R^2 index in the FE model, following the recommendations of Konstantopoulos and Hedges (2009). Equations used for moderator analyses are presented in Table 3. Negative values in R^2 were truncated to zero.

From the statistical theory, some expectations arise. Firstly, statistically significant relationships are more likely when the statistical analyses are based on the FE model,

Table 4. Descriptive statistics of the five RG examples

	RG example				
Statistic	*MOCI*	*STAI*	*Y-BOCS*	*HAM-D*	*MBI*
k	39	29	40	25	51
$r_{\alpha n}$	−0.186	0.337	0.269	0.291	0.050
p	.257	.074	.094	.158	.726
Min.	0.610	0.790	0.690	0.420	0.660
Max.	0.870	0.940	0.960	0.897	0.950
Mean	0.767	0.891	0.848	0.747	0.870
Median	0.760	0.898	0.880	0.780	0.880
SD	0.060	0.034	0.072	0.133	0.049
Skewness	−0.348	−0.955	−0.906	−1.170	−1.784
Kurtosis	0.059	1.631	−0.132	0.937	5.916
S-W test:					
Alpha	.538	0.035	< .001	.001	< .001
Fisher's *z*	.656	0.634	.034	.157	.282
Hakstian–Whalen	.741	0.419	.013	.060	.051
Bonett	.557	0.666	.039	.284	.403

Note. *k*: number of studies. $r_{\alpha n}$: Pearson correlation between sample size and coefficient alpha. *p*: probability value of the statistical test for $r_{\alpha n}$. Min. Max.: Minimum and maximum coefficients alpha. *SD*: Standard deviation. S-W test: *p*-value obtained by the Shapiro–Wilk test for testing normality in the data set, taking untransformed coefficients alpha (alpha) or using the Fisher *z*, Hakstian–Whalen, or Bonett transformations.

followed by the VC model, the RE models, and OLS methods. This is because the FE model assumes a common coefficient alpha and, as a consequence, takes into account only one source of variability: the within-study variance. The VC model will produce more conservative results than the FE model because it assumes heterogeneity among the true coefficients alpha. In turn, the RE models tend to produce more conservative results than the VC model because they assume a superpopulation of true coefficients alpha and, as a consequence, take into account two sources of variability: the within-study and between-studies variability. Finally, the OLS method is usually the most conservative analysis because it does not take advantage of accumulating the sample sizes of the studies (Bonett, 2010; Borenstein, Hedges, Higgins & Rothstein, 2009, 2010; Rodriguez & Maeda, 2006). Secondly, only when there is a strong or a null relationship between reliability estimates and a moderator variable should the different statistical methods coincide in their results. Thirdly, when there is a moderate relationship, the results obtained from the different statistical methods may differ in terms of statistical significance and proportion of variance accounted for.

In order to describe the shape of the distribution of the coefficients alpha, Figure 1 presents the histograms of the five RG examples, and Table 4 reports the main descriptive statistics for each database. With the exception of the MOCI example, histograms exhibited a pronounced skewness and the Shapiro–Wilk tests for testing normality were statistically significant (Table 4). All transformations, and particularly Bonett's, provided a better fit to normality. However, the lack of statistical power of the Shapiro–Wilk test with such a small number of studies as in our five examples necessitates a cautious interpretation of the results (cf. Keskin, 2006).

Table 5. Comparison of results in terms of average reliability coefficient and confidence interval width

Statistical model	Transformation method	MOCI			STAI			Y-BOCS			HAM-D			MBI		
		$\bar{\alpha}$	*D* (%)	Width	$\bar{\alpha}$	*D* (%)	Width	$\bar{\alpha}$	*D* (%)	Width	$\bar{\alpha}$	*D* (%)	Width	$\bar{\alpha}$	*D* (%)	Width
OLS	Untransformed	0.767	–	0.039	0.891	–	0.026	0.848	–	0.046	0.747	–	0.110	0.870	–	0.028
OLS	Fisher's *z*	0.774	0.86	0.039	0.895	0.49	0.024	0.862	1.61	0.041	0.771	3.14	0.093	0.877	0.80	0.024
FE	Fisher's *z*	0.759	−1.03	0.016	0.904	1.46	0.008	0.880	3.75	0.012	0.810	8.45	0.018	0.878	0.93	0.006
FE	Hakstian–Whalen	0.764	−0.35	0.013	0.906	1.77	0.006	0.887	4.51	0.009	0.833	11.50	0.013	0.882	1.41	0.004
FE	Bonett	0.760	−0.94	0.013	0.904	1.48	0.006	0.880	3.79	0.010	0.814	8.87	0.014	0.878	0.96	0.005
VC	Untransformed	0.767	–	0.019	0.891	–	0.012	0.848	–	0.032	0.747	–	0.033	0.870	–	0.008
RE	Fisher's *z*	0.767	−0.06	0.032	0.898	0.85	0.020	0.867	2.19	0.032	0.771	3.16	0.100	0.876	0.77	0.019
RE	Hakstian–Whalen	0.770	0.37	0.031	0.898	0.78	0.019	0.868	2.28	0.029	0.769	2.90	0.083	0.876	0.72	0.018
RE	Bonett	0.769	0.24	0.031	0.897	0.76	0.019	0.866	2.14	0.031	0.775	3.66	0.095	0.877	0.83	0.019
RE-c	Fisher's *z*	0.767	−0.06	0.036	0.898	0.85	0.022	0.867	2.19	0.038	0.771	3.16	0.094	0.876	0.77	0.023
RE-c	Hakstian–Whalen	0.770	0.37	0.037	0.898	0.78	0.023	0.868	2.28	0.038	0.769	2.90	0.095	0.876	0.72	0.024
RE-c	Bonett	0.769	0.24	0.037	0.897	0.76	0.023	0.866	2.14	0.038	0.775	3.66	0.089	0.877	0.83	0.023
RE-n	Untransformed	0.755	−1.51	0.030	0.901	1.14	0.019	0.872	2.85	0.032	0.784	4.93	0.106	0.874	0.49	0.019

Note. $\bar{\alpha}$: average reliability coefficient (back-transformed where necessary). *D*: discrepancy index. Width: width for the confidence interval ($1 - \alpha = 0.95$).

Table 6. Comparison of results regarding moderator analyses

		MOCI				STAI				Y-BOCS				HAM-D				MBI			
		Test version		*SD*		Age		*SD*		Test version		*SD*		Disorder		*SD*		Test version		*SD*	
Mod.	Trn.	*p*	R^2	*p*	R^2	*p*	R^2	*p*	R^2	*p*	R^2	*p*	R^2	*p*	R^2	*p*	R^2	*p*	R^2	*p*	R^2
OLS	–	.044	.081	.000	.516	.013	.178	.000	.440	.481	0	.241	.014	.442	0	.059	.190	.098	.036	.000	.524
OLS	*Z*	.035	.091	.000	.521	.026	.140	.000	.430	.325	0	.435	0	.318	.002	.132	.101	.066	.049	.000	.437
FE	*Z*	.000	.075	.000	.697	.000	.150	.000	.395	.349	0	.023	.073	.000	.224	.000	0	.000	0	.000	.009
FE	*T*	.000	.084	.000	.623	.000	.103	.000	.330	.240	0	.154	0	.000	.210	.000	0	.000	0	.000	0
FE	*L*	.000	.067	.000	.617	.000	.132	.000	.351	.303	0	.005	.029	.000	.221	.000	0	.000	0	.000	0
VC	*L*	.000	.092	.000	.520	.000	.138	.000	.428	.068	.001	.166	0	.003	.006	.000	.084	.000	.050	.000	.426
RE	*Z*	.044	.075	.000	.697	.009	.150	.000	.395	.621	0	.395	.073	.238	.224	.257	0	.048	0	.000	.009
RE	*T*	.026	.084	.000	.623	.007	.103	.000	.330	.478	0	.421	0	.179	.210	.192	0	.033	0	.000	0
RE	*L*	.029	.067	.000	.617	.007	.132	.000	.351	.469	0	.414	.029	.239	.221	.284	0	.037	0	.000	0
RE-c	*Z*	.068	.075	.000	.697	.019	.150	.000	.395	.672	0	.455	.073	.248	.224	.141	0	.099	0	.000	.009
RE-c	*T*	.054	.084	.000	.623	.023	.103	.000	.330	.579	0	.496	0	.273	.210	.125	0	.086	0	.000	0
RE-c	*L*	.060	.067	.000	.617	.020	.132	.000	.351	.549	0	.486	.029	.248	.221	.161	0	.086	0	.000	0
RE-n	–	.044	.081	.000	.516	.013	.178	.000	.440	.481	0	.241	.014	.442	0	.059	.190	.098	.036	.000	.524

Note. Mod.: Statistical model. Trn.: Transformation method for the coefficients (*Z,* Fisher's *z*; *T*, Hakstian–Whalen*; L*, Bonett). *SD*: standard deviation of the test scores. *p*: *p*-value for each statistical test. For *t*-tests, the *p*-value was obtained from a *t*-distribution with $k - 2$ degrees of freedom. R^2: proportion of variance explained by the moderator in the simple regression analyses (when negative, it is truncated to zero).

3.1. Average reliability estimate

The first purpose of this study was to compare the average reliability coefficient and the confidence interval obtained with each of the 13 statistical procedures described above. Table 5 presents the results for the five RG examples.

Taking the unweighted mean of untransformed coefficients as reference, most of the average estimates showed discrepancies smaller than 5%. OLS and RE methods provided lower discrepancies, while FE methods led to the highest percentages in most situations. More specifically, combining FE methods and Hakstian–Whalen transformation resulted in the highest discrepancy index in four of the five examples, including the maximum discrepancy rate, 11.50% for the HAM-D scale. It is worth noting that the highest discrepancies were achieved for the Y-BOCS and HAM-D examples, whose coefficient distributions are skewed even after applying transformations (Table 4).

General trends when comparing confidence interval widths indicate a clear influence of the statistical model, rather than the transformation method. Accordingly, OLS methods provided the widest intervals in most situations, followed by RE-c methods, RE models (including RE-n), the VC model and, lastly, FE methods, which showed narrower intervals in all examples analysed here.

3.2. Searching for moderator variables

For illustrative purposes, in each example we focused on one dichotomous and one continuous predictor, in order to examine their possible association with reliability coefficients. In the five examples the continuous predictor was the standard deviation of test scores. The test version (original vs. adapted to other languages) was the dichotomous predictor in the MOCI, Y-BOCS, and MBI examples, whereas the mean age ($\leq$ 65 vs. $>$ 65 years) and the type of disorder (depression vs. other) were the dichotomous predictors in the STAI and HAM-D examples, respectively. Table 6 presents the results of the moderator analyses, reporting p-values and R^2 indices for each linear regression model fitted.

Significance tests, represented by p-values, showed agreement across methods when there was a strong relationship of the moderator to the coefficients (e.g., standard deviation in the MOCI and MBI examples), or when that relationship was almost null, as seen for the test version in the Y-BOCS example. However, disagreement arose in the statistical conclusion when the relationships were not that extreme, and the most conservative results were provided by RE-c methods. Conversely, the lowest p-values were reported by FE methods, followed by the VC model.

For the R^2 index, comparisons are restricted to two main indices (Table 3): the adjusted R^2 employed in OLS, VC, and RE-n methods and the meta-analytic version of the R^2 index proposed by Raudenbush (2009). For most situations, the former procedure provided higher values. An unexpected finding was that, for some examples (standard deviation in HAM-D and MBI scales, test version in the MBI scale), a statistically significant result corresponded to a percentage of variance accounted for close to zero, particularly for WLS methods. These results suggest a larger statistical power for weighted procedures.

4. Discussion and conclusions

This study has examined the extent to which different statistical procedures applied in RG studies lead to different results and conclusions when calculating an average reliability

coefficient with its confidence interval and looking for possible moderators of reliability. To this end, statistical methods based on different statistical models for integrating coefficients alpha were applied to the data from five real RG studies and their results were compared. The main discrepancies among the different statistical models refer to whether or not to transform the coefficients and how to weight them. Thirteen different statistical methods were compared.

Regarding the average reliability coefficient, the different statistical procedures compared here seem to give, in general, similar results although, depending on the database, important differences can be found. It seems that the most important factors that lead to differences in the overall reliability estimates are the weighting method and the shape of the distribution of the coefficients alpha.

The width of the confidence interval computed around the average reliability coefficient seems to be affected by the procedure selected. As expected, FE methods produced narrower intervals, followed by the VC model and, lastly, RE and OLS methods. This is due to differences in the error sources taken into account by the alternative statistical models: (1) FE models take into account the within-study variability as the only error source; (2) the VC model also considers within-study variability, but assumes that the studies are estimating different population coefficients; and (3) RE models take into account both within-study and between-studies variability, as they estimate the mean of a potential population of coefficients alpha larger than the set of studies included in the meta-analysis. The RE-c model also accounts for the uncertainty of estimating sampling variances, providing wider confidence intervals.

The main differences among the statistical models were found in the analysis of moderator variables. Obviously, when there was a strong or a null relationship between the moderator variable and reliability estimates, all methods coincided in their results. Differences among methods were especially notable when the association between the predictor variable and reliability coefficients was of small to medium magnitude. As expected, the FE model reached statistical significance most often, due to the aforementioned reasons. In some cases the transformation method of coefficients alpha also affected the results from this model, as in the analysis of the standard deviation from the Y-BOCS, where statistical significance was reached using the Fisher *z*- and Bonett transformations, but not using Hakstian–Whalen.

From the psychometric theory a positive relationship was expected between the test score variability and reliability. This expectation was not always borne out. Out of our five examples, we found a positive, significant relationship between the standard deviation of test scores and coefficient alpha in the MOCI, STAI, and MBI studies, with virtually all statistical models. Conversely, this relationship was not so evident across the different methods for the Y-BOCS and HAM-D examples. Therefore, depending on the data set and the statistical model assumed, a positive relationship between reliability and test score variability may or may not be found.

4.1. Which model should we use?

Our study showed that applying different statistical models to the same data from an RG study can affect the results and conclusions reached. At the same time, RG studies that have applied different statistical models can also give results that are not comparable to each other. Next, we present some guidelines on how a meta-analyst should select the statistical model to conduct an RG meta-analysis with coefficients alpha. These recommendations are not based on empirical comparison among the different methods

presented here, but on methodological work and simulation studies carried out previously both in the more general meta-analytic field and in the more specific RG approach.

There is a general consensus to consider that the main criterion for choosing the statistical model in a meta-analysis should be the extent to which the meta-analyst aims to generalize his/her results (Borenstein *et al.*, 2009, 2010; Field & Gillett, 2010; Hedges & Vevea, 1998; Overton, 1998; Schmidt, 2010; Schmidt, Oh, & Hayes, 2009). If the meta-analyst intends to generalize results to a population of studies identical to those included in the meta-analysis, then FE and VC are appropriate models. The latter seems more realistic since, unlike the former, it assumes that each study estimates a different population coefficient. Thus, depending on the heterogeneity exhibited by the reliability coefficients, the meta-analyst must select an FE or a VC model. The value of the I^2 index can be used to assess whether the set of reliability coefficients is reasonably homogeneous or not. In this respect, a guiding benchmark is to assume an FE model when I^2 is under 25% (Higgins & Thompson, 2002), and a VC model otherwise.

Due to the large heterogeneity usually exhibited by the reliability coefficients obtained in different applications of the same test, the potential for assuming an FE model is very limited. Homogeneity will occur in exceptional cases – for example, a small number of reliability coefficients (about 10 or 15) obtained from samples selected from the same population (participants with a given psychological disorder, similar age, gender distribution, etc.) and under similar conditions and settings (samples of inpatients, selected from English-speaking countries, etc.). In these cases, it will be reasonable to assume that all of the reliability coefficients are estimating a common population reliability coefficient and, consequently, the main purpose of the RG study should be to estimate the common coefficient alpha and to make generalizations to a population of studies with characteristics similar to those of the studies in the meta-analysis (Borenstein *et al.*, 2010). Our comparative results as well as those of recent simulation work (López-Pina, Sánchez-Meca, & López-López, 2012) indicate that, under the homogeneity assumption, whether or not we transform the coefficients barely has an influence on the average coefficient alpha, but the weighted methods are more efficient than the unweighted ones.

When the set of coefficients alpha exhibits a heterogeneity that cannot be explained only by random sampling of the participants in the samples, then the FE model is not acceptable and the meta-analyst will have to decide between the VC and RE models. In this case, the main purpose of the meta-analysis will be to identify study characteristics that can explain at least part of the variability exhibited by the reliability estimates, although a mean reliability coefficient will also be of interest as a summary result of the test scores' reliability. In choosing between the two models, it is important to note that RE models need more assumptions to be met than the VC model.

Selecting a VC model can be justified for several reasons. Firstly, if the meta-analyst intends to generalize his/her results to a population of studies with characteristics similar to those of the studies included in the RG study, then a VC model will be the best option. Secondly, if the conditions for the RE model are not met, then the meta-analyst must assume the VC model, even if he/she originally intended to apply an RE model. For example, if the number of coefficients alpha (or studies) is small, then applying an RE model will be very risky, as the between-studies variance will be very poorly estimated (Bonett, 2010; Borenstein *et al.*, 2009) and, as a consequence, the VC model will be a better option than the RE model. In fact, Bonett (2010) proposed using the VC model for a small number of very carefully selected reliability coefficients: 'Here the recommendation is to use only carefully selected and high-quality studies for the specific purpose of

obtaining a more accurate average reliability estimate of a specific measurement scale, generalizing the reliability results to several clearly defined study populations, or to assess the effects of specific moderator variables on the reliability of a measurement scale' (p. 380). Some guidelines on how many studies are needed to obtain a reasonably accurate estimate of the between-studies variance are outlined later.

More often, however, generalization is intended to a larger population of studies than those included in the meta-analysis. The aim in meta-analysis, as in any research, is to generalize the results beyond the specific units used: 'The usual goal of research … is generalizable knowledge …, which requires generalization beyond the current set of studies to other similar studies that have been or might be conducted' (Schmidt *et al.*, 2009, p. 101). Consequently, RE models are conceptually more appropriate for most situations when conducting a meta-analysis (Borenstein *et al.*, 2010; Field, 2003, 2005; National Research Council, 1992; Schmidt, 2010).

However, the application of RE models entails three main problems. Firstly, in practice the studies in a meta-analysis are not randomly selected from a larger population of studies and therefore, in the strictest sense, it is not appropriate to estimate the average population coefficient of this superpopulation of potential studies. This is a criticism raised by Bonett (2008, 2009, 2010) against the use of RE models. Secondly, with a small number of studies, estimates of the between-studies variance are very inaccurate, which will also affect the statistical analyses conducted with RE models. Thirdly, if the normality assumption of the hypothetical superpopulation of reliability coefficients cannot be maintained, then statistical inferences from the RE model may be inadequate.

Regarding the first problem, as stated by Laird and Mosteller (1990, p. 14), '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'. Therefore, if this criticism is extended to primary studies, then no meta-analytic model would be appropriate, since the vast majority of primary studies strictly violate the random sampling assumption. However, statistical inference techniques are applied routinely in primary research, and primary researchers routinely generalize their results to a population of units. Likewise, the meta-analyst will apply RE models when he/she can assume, on a reasonable basis, that the set of studies included in the meta-analysis is a representative sample of a potential population of past or future studies. To apply RE models, the meta-analyst must define, also on a reasonable basis, the characteristics of the potential population of studies to which he/she aims to generalize the results. In the RG context, the target population would be the set of studies using a specific measurement instrument, regardless of whether the studies provided a reliability estimate or not. This target population will be appropriate because, *a priori*, there is no reason to believe that studies that reported reliability are different in composition and variability to those that did not report it. But in any case, it is easy to check this by comparing statistically the substantive and methodological characteristics of both groups of studies.

Other problems when applying RE models refer to the difficulties in accurately estimating the between-studies variance when the number of studies is small and the normality assumption of the true reliability coefficient distribution is not met. Based on Field's (2005) simulation work, Aguinis, Gottfredson, and Wright (2011, p. 1039) stated: 'The RE methods require at least 20 primary-level studies in order to obtain a properly performing [confidence interval] for the mean effect size assuming approximate normality of the superpopulation of effect sizes'. Similar recommendations were proposed by Biggerstaff and Tweedie (1997) and Brockwell and Gordon (2001). On the other hand, based on several simulation studies, Schulze (2004) concluded that at least 32

studies will be needed to apply mixed-effects meta-regression models to warrant a reasonably good performance of the parameter estimators (in particular, the residual between-studies variance). However, based on his simulation results, Bonett (2010) is much more restrictive in applying RE models: 'a large number of studies ($k > 150$) are needed to detect the degree of kurtosis that would cause serious problems with a confidence interval for τ, implying that RE meta-analyses should not be attempted with fewer than 150 studies' (p. 371). These recommendations refer to the conventional RE model, not to the RE-c model. Recent simulation work, including non-normality for the true reliability coefficients' distribution, has shown better performance for the RE-c model than for the conventional RE methods when mixed-effects meta-regression models are conducted, in terms of the efficiency of the parameter estimators, adjustment to the nominal confidence level, and statistical power (López-López, Botella, Sánchez-Meca, & Marín-Martínez, in press; Viechtbauer, López-López, Sánchez-Meca, & Marín-Martínez, 2012). Based on recent simulation work, our recommendation is that with 30 or more studies (or reliability estimates), RE methods will perform reasonably well, even when the normality assumption is not strictly met. However, if the distribution of coefficients departs very significantly from normality, even once the coefficients alpha are transformed, more than 30 studies will be needed to guarantee good performance by the RE model. With a small number of studies, we recommend generalizing the results only to a population of studies identical in composition and variability to those included in the meta-analysis and, therefore, assuming a VC model.

In summary, since the goal of meta-analysis is to generalize knowledge, RE models are optimal in most applications of the RG approach. Therefore, when the assumptions of the RE model are reasonably met, our recommendation is to apply Bonett's transformation on coefficients alpha for the statistical analyses, as well as the corrections proposed by Hartung (1999) and Knapp and Hartung (2003; i.e., the RE-c model), given their good performance in previous simulations. It is worth noting that although we have focused on the RG approach for coefficient alpha, most of the arguments presented here can be generalized for use in RG studies that integrate other types of reliability coefficients, such as test–retest, parallel forms or intra-class correlations. Note that the Hakstian–Whalen and Bonett transformations should not be used in such cases, as they have been specifically proposed for transforming coefficients alpha. Finally, more simulation work is required in order to examine the performance of the different statistical models under a wide range of conditions and violations of the assumptions such as that of normality.

Acknowledgement

This research was supported by a grant from the Fundación Séneca, Region of Murcia (Spain; Project No. 08650/PHCS/08).

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. doi:10.1002/job.719

Aloe, A. M., Becker, B. J., & Pigott, T. D. (2010). An alternative to R^2 for assessing linear models of effect size. *Research Synthesis Methods*, *1*, 272–283. doi:10.1002/jrsm.23

Bachner, Y. G., & O'Rourke, N. (2007). Reliability generalization of responses by care providers to the Zarit Burden Interview. *Aging and Mental Health*, *11*, 678–685. doi:10.1080/13607860701529965

Beretvas, S. N., & Pastor, D. A. (2003). Using mixed-effects models in reliability generalization studies. *Educational and Psychological Measurement*, *63*, 75–95. doi:10.1177/0013164402239318

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. doi:10.1177/0013164407301529

Biggerstaff, B. J., & Tweedie, R. L. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine*, *16*, 753–768. doi:10.1002/(SICI)1097-0258(19970415)16:73.3.CO;2-7

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*, 335–340. doi:10.3102/10769986027004335

Bonett, D. G. (2008). Meta-analytic interval estimation for bivariate correlations. *Psychological Methods*, *13*, 173–181. doi:10.1037/a0012868

Bonett, D. G. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods*, *14*, 225–238. doi:10.1037/a0016619

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods*, *15*, 368–385. doi:10.1037/a0020142

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

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. doi:10.1002/jrsm.12

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 variance in studies of internal consistency generalization. *Methodology*, *8*(2), 71–80. doi:10.1027/1614-2241/a000039

Botella, J., Suero, M., & Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods*, *15*, 386–397. doi:10.1037/a0019626

Brockwell, S. E., & Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, *20*, 825–840. doi:10.1002/sim.650

Campbell, J. S., Pulos, S., Hogan, M., & Murry, F. (2005). Reliability generalization of the Psychopathy Checklist applied in youthful samples. *Educational and Psychological Measurement*, *65*, 639–656. doi:10.1177/0013164405275666

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

DerSimonian, R., & Kacker, R. (2007). Random-effects models for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials*, *28*, 105–114. doi:10.1016/j.cct.2006.04.004

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

Feldt, L. S., & Charter, R. A. (2006). Averaging internal consistency reliability coefficients. *Educational and Psychological Measurement*, *66*, 215–227. doi:10.1177/0013164404273947

Field, A. P. (2003). The problems of using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics*, *2*, 77–96. doi:101207/S15328031US0202_02

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods*, *10*, 444–467. doi:10.1037/1082-989X.10.4.444

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*, 665–694. doi:10.1348/000711010X502733

Graham, J. M., & Christiansen, K. (2009). The reliability of romantic love: A reliability generalization meta-analysis. *Personal Relationships*, *16*, 49–66. doi:10.1111/j.1475-6811.2009.01209.x

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

Hakstian, A. R., & Whalen, T. E. (1976). A *k*-sample significance test for independent alpha coefficients. *Psychometrika*, *41*, 219–231. doi:10.1007/BF02291840

Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal*, *41*, 901–916. doi:10.1002/(SICI)1521-4036(199912)41:83.0.CO;2-W

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

Hedges, L. V., & Olkin, I. (1985). *Statistical methods in 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. doi:10.1037//1082-989X.3.4.486

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.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539–1558. doi:10.1002/sim.1186

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

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

Keskin, S. (2006). Comparison of several univariate normality tests regarding Type I error rate and power of the test in simulation based small tests. *Journal of Applied Science Research*, *2*, 296–300.

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: Russell Sage Foundation.

Laird, N. M., & Mosteller, F. (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care*, *6*, 5–30. doi:10.1017/S0266462300008916

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. doi:10.1177/0013164405284030

López-López, J. A., Botella, J., Sánchez-Meca, J., & Marín-Martínez, F. (in press). Alternatives for mixed-effects meta-regression models in the reliability generalization approach: A simulation study. *Journal of Educational and Behavioral Statistics*.

López-Pina, J. A., Sánchez-Meca, J., & López-López, J. A. (2012). Methods for averaging alpha coefficients in reliability generalization studies. *Psicothema*, *24*, 161–166.

López-Pina, J. A., Sánchez-Meca, J., Marín-Martínez, F., Núñez-Núñez, R. M., Rosa-Alcázar, A. I., Gómez-Conesa, A., & López-López, J. A. (2010, July). *La Escala de Obsesiones y Compulsiones Yale-Brown (Y-BOCS): Un estudio de generalización de la fiabilidad*. [The Yale-Brown Obsessive-Compulsive Scale (Y-BOCS): A reliability generalization study.] Paper presented at the VII Congreso Iberoamericano de Psicología, Oviedo, Spain.

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.

Mason, C., Allam, R., & Brannick, M. T. (2007). How to meta-analyze coefficient-of-stability estimates: Some recommendations on Monte Carlo studies. *Educational and Psychological Measurement*, *67*, 765–783. doi:10.1177/0013164407301532

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

O'Rourke, N. (2004). Reliability generalization of responses by care providers to the Center for Epidemiologic Studies-Depression Scale. *Educational and Psychological Measurement*, *64*, 973–990. doi:10.1177/0013164404268668

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. doi:10.1037//1082-989X.3.3.354

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

Rendina-Gobioff, G. (2004).*A meta-analysis reliability generalization study: Reliability of the Statistical Anxiety Rating Scale (STARS) Subscale* Worth of Statistics. Unpublished manuscript, University of South Florida.

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, *11*, 306–322. doi:10.1037/1082-989X.11.3.306

Romano, J. L, Kromrey, J. D., & Hibbard, S. T. (2010). A Monte Carlo study of eight confidence interval methods for coefficient alpha. *Educational and Psychological Measurement*, *70*, 376–393. doi:10.1177/0013164409355690

Rouse, S. V. (2007). Using reliability generalization methods to explore measurement error: An illustration using the MMPI-2 PSY-5 scales. *Journal of Personality Assessment*, *88*, 264–275. doi:10.108/00223890701293908

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. doi:10.1037/1082-989X.13.1.31

Sawilowsky, S. 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. doi:10.1177/00131640021970439

Schmidt, F. L. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science*, *5*, 233–242. doi:10.1177/174569161036933

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. doi:10.1348/000711007X255327

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

Sidik, K., & Jonkman, J. N. (2005). A note on variance estimation in random effects meta-regression. *Journal of Biopharmaceutical Statistics*, *15*, 823–838. doi:10.1081/BIP-200067915

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*, 174–195. doi:10.1177/0013164400602002

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6–20. doi:10.1177/0013164498058001002

Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, *62*, 562–569. doi:10.1177/001316402128775012

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. doi:10.1177/0748175611409845

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. doi:10.1007/s11136-008-9398-2

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-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48.

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., & Marín-Martínez, F. (2012). *A comparison of procedures to test for moderators in meta-regression models*. Unpublished manuscript.

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement*, *60*, 201–223. doi:10.1177/00131640021970466

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. doi:10.1891/jnum.2005.13.1.7

Received 07 November 2011; revised version received 28 April 2012