Current Psychology (2024) 43:18275–18293
https://doi.org/10.1007/s12144-023-05604-y

Reliability generalization meta-analysis: comparing different statistical methods

Carmen López-Ibáñez[1] · Rubén López-Nicolás[1] · Desirée M. Blázquez-Rincón[1,2] · Julio Sánchez-Meca[1]

Accepted: 22 December 2023 / Published online: 22 January 2024

Abstract

Reliability generalization (RG) is a kind of meta-analysis that aims to characterize how reliability varies from one test application to the next. A wide variety of statistical methods have typically been applied in RG meta-analyses, regarding statistical model (ordinary least squares, fixed-effect, random effects, varying-coefficient models), weighting scheme (inverse variance, sample size, not weighting), and transformation method (raw, Fisher's Z, Hakstian and Whalen's and Bonett's transformation) of reliability coefficients. This variety of methods compromise the comparability of RG meta-analyses results and their reproducibility. With the purpose of examining the influence of the different statistical methods applied, a methodological review was conducted on 138 published RG meta-analyses of psychological tests, amounting to a total of 4,350 internal consistency coefficients. Among all combinations of procedures that made theoretical sense, we compared thirteen strategies for calculating the average coefficient, eighteen for calculating the confidence intervals of the average coefficient and calculated the heterogeneity indices for the different transformations of the coefficients. Our findings showed that transformation methods of the reliability coefficients improved the normality adjustment of the coefficient distribution. Regarding the average reliability coefficient and the width of confidence intervals, clear differences among methods were found. The largest discrepancies were found between the different strategies for calculating confidence intervals. Our findings point towards the need for the meta-analyst to justify the statistical model assumed, as well as the transformation method of the reliability coefficients and the weighting scheme.

Keywords Meta-analysis · Reliability generalization · Statistical models · Reliability coefficient

Meta-analysis has become an essential method to integrate the results of studies that address a given question. Typical meta-analyses in psychology aim to answer such questions as the efficacy of interventions, to identify risk or protection factors to suffer a given problem, or to estimate the magnitude of associations between variables. To accomplish these objectives, meta-analyses use such effect size indices as standardized mean differences, correlation coefficients, or odds ratios. In the last 20 years, research in psychology and other related health sciences has dedicated some attention to a special kind of meta-analysis usually named 'reliability generalization meta-analysis'. The term reliability generalization (RG) was coined by Vacha-Haase (1998) to refer to a meta-analysis aimed at characterizing how measurement error of the test scores varies as applied from one sample to the next. Unlike typical meta-analyses, in an RG meta-analysis the 'effect size' is the reliability coefficient reported in studies that have applied a given measurement tool, such as alpha coefficients, test–retest, parallel-form, or inter-rater coefficients, among others (Botella et al., 2010; Henson & Thompson, 2002; Mason et al., 2007; Sánchez-Meca et al., 2021; Thompson, 2003; Vacha-Haase et al., 2002).

The rationale for this kind of meta-analysis is that, as classical test theory states, reliability is not an inherent property of the test, but varies as the test is applied to different samples (Crocker & Algina, 1986; Gronlund & Linn, 1990; Traub, 1994). Sentences like 'the test reliability is 0.8' are incorrect, as they assume that reliability is an immutable property of the test. It is more appropriate

✉ Carmen López-Ibáñez
carmen.li@um.es
https://www.um.es/metaanalysis

[1] Department of Basic Psychology and Methodology, Faculty of Psychology, University of Murcia, Murcia, Spain

[2] Department of Psychology and Education, Faculty of Health Sciences and Education, Madrid Open University, Madrid, Spain

to say that 'the reliability of the test scores in this sample is 0.8'. Reliability of test scores is one of the most important psychometric properties of a measurement tool, such that it is important to investigate how reliability varies as applied in different samples and which study characteristics can explain this variability. These questions are relevant, regardless of the type of reliability investigated: internal consistency (Cronbach's alpha, parallel-forms, omega coefficients), temporal stability (retest correlations), or inter-rater agreement (inter-rater coefficients such as Cohen's kappa, intraclass correlations). It is important to note that studies that induce the reliability of test scores, that is, that endorse reliability estimates from other previous studies, cannot be included in an RG meta-analysis. Only genuine reliability estimates with the data at hand of the primary studies can be included in an RG meta-analysis (Sánchez-Meca et al., 2021).

As reliability varies from one test application to the next, meta-analysis is an optimal methodology to investigate which study characteristics can be statistically associated to the reliability estimates variability. Thus, by integrating all single studies that have applied a given test and have reported a reliability estimate with data at hand, an RG meta-analysis enables us to estimate the average reliability of test scores and to identify potential moderator variables of that reliability. Study characteristics that may affect reliability estimates are composition and sample variability, target population to which the sample of participants pertain (e.g., community, subclinical, clinical population), the test adaptation to different languages and cultures, or the context where the test is applied (Botella & Ponte, 2011; Botella & Suero, 2012; Henson & Thompson, 2002; Rodriguez & Maeda, 2006; Thompson, 2003).

To date, a large number of RG meta-analyses have been carried out in psychology on different measurement instruments. A systematic search has identified more than 150 RG meta-analyses conducted on psychological measurement tools between 1998 and 2019 (Sánchez-Meca et al., 2019). Sánchez-Meca et al. (2013) identified a variety of methods to statistically integrate reliability coefficients. Differences among the methods refer to the statistical model assumed (e.g., fixed-effect versus random-effects models), whether reliability estimates must be transformed to normalize their distribution and stabilize their variances, and whether to weight the reliability estimates when they are statistically integrated. An issue not yet investigated is whether the choice of different statistical methods can lead to substantial changes in RG meta-analysis results. If different methods applied to the same RG meta-analysis have an impact in their results, then their conclusions will be affected by the methods applied. In addition, the results of RG meta-analyses applying different methods cannot be compared.

Purpose

Applying different statistical models and methods to synthesize a set of reliability coefficients on a given test can lead to different findings, affecting their conclusions. To our knowledge, attempts to investigate this problem have not yet been accomplished. The main purpose of this research was to examine the extent to which different statistical methods to obtain a pooled reliability coefficient and a confidence interval around it can lead to different results. With this aim, a methodological review was conducted of all RG meta-analyses on psychological tools published to date. An exhaustive search was performed to identify RG meta-analyses carried out on psychological scales, to obtain their datasets, to apply different statistical methods, and to compare their results. It is worth noting that this investigation is not a simulation study aimed at determining which statistical methods exhibit better properties. This study is an empirical comparison of alternative statistical methods for conducting an RG meta-analysis in order to examine the extent to which different methods can affect the meta-analytic results.

As internal consistency is the most frequently reported type of reliability in RG meta-analyses, our study focused on internal consistency coefficients such as Cronbach's alpha, parallel-form, or omega coefficients. In particular, we tried to ascertain the extent to which different methods to average a set of internal consistency reliability coefficients provide heterogeneous results depending on whether to transform reliability coefficients, the statistical model assumed, and the weighting factor applied. In addition, we also aimed to compare different methods to construct a confidence interval for the average reliability coefficient, as regards confidence width. Another purpose consisted of examining the extent to which different transformation methods devised to normalize reliability coefficient distribution achieve this objective. Finally, we also wished to compare the amount of heterogeneity (quantified with the I^2 index and prediction intervals) exhibited by untransformed and transformed reliability coefficients. In the next sections, different methods to statistically integrate reliability coefficients are presented and the methodology of this meta-review is outlined. Findings comparing the results of the different methods applied to the RG meta-analyses are then described, and finally the scope of our results is discussed.

In recent years, the reproducibility and replicability of psychological research have become important topics (McNutt, 2014; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). Meta-analyses are not immune to these issues; therefore, efforts to investigate factors that may affect the reproducibility of meta-analyses are warranted (Lakens et al., 2016). It is important to note that the purpose of this investigation was not to reproduce the original findings reported in the RG meta-analyses but to examine the extent to which different statistical methods applied to the same meta-analytic datasets can lead to different results.

Statistical methods in RG meta-analysis

When conducting an RG meta-analysis, the meta-analyst must decide whether reliability coefficients should be transformed to normalize distribution and stabilize variances. Some authors advise against transforming reliability coefficients (Henson & Thompson, 2002; Mason et al., 2007), whereas others are in favor (Rodriguez & Maeda, 2006; Sánchez-Meca et al., 2013). Thus, some RG meta-analyses did not transform internal consistency coefficients, whereas others applied some of the following transformations: Fisher's *Z*, Hakstian and Whalen's transformation (1976), and Bonett's one (2002). Not all transformations are recommended for all types of reliability coefficients: for example, for indices based on Pearson's correlation coefficient such as split-half reliability coefficients, the most appropriate transformation would be Fisher's Z transformation. However, for those coefficients ranging between 0 and 1 (Cronbach's alpha, McDonald's omega or split-rater reliability, among others), it would be theoretically more correct to use Bonett's (2002) transformation.

In this meta-review we have selected those transformations that have been most frequently applied in RG meta-analyses. Table 1 presents the different methods to transform internal consistency coefficients with their corresponding sampling variances, as well as formulas to back-transform the transformed coefficients to the original metric (Sánchez-Meca et al., 2013). Note that in Table 1 the typical symbol to represent Cronbach's alpha reliability coefficients is used ($\widehat{\alpha}_i$), as most RG meta-analyses use this coefficient to estimate the internal consistency of scales. This is due to alpha coefficients being routinely reported in primary studies. However, formulas shown in Table 1 can also be applied to other types of internal consistency reliability coefficients.

Another important decision in a meta-analysis is choosing the statistical model under which the statistical analyses will be accomplished. Fixed-effect (FE), in singular, and random-effects (RE) models are the two most commonly used statistical models in meta-analysis. Under an FE model (also named *common-effect model*) the meta-analyst assumes that the reliability coefficients reported in the studies are estimating a common population parameter, so that the only variability source among reliability estimates is due to sampling error. When an RE model is assumed, the meta-analyst is then acknowledging that the reliability estimates exhibit more variability than sampling error can explain. The extra heterogeneity is due to the fact that each reliability coefficient is estimating a different parameter, these parameters constituting a representative sample of a distribution of potential parametric reliability coefficients. RE model takes into account two variability sources: within-study variance (i.e., the same as in the FE model), due to sampling of participants in each sample, and between-studies variance, owing to sampling of true reliability coefficients from a super-population of reliability coefficients. Assuming one or another statistical model has consequences on how statistical analyses are accomplished and on the degree of generalizability of the meta-analytic results. In particular, how the reliability estimates are weighted is different depending on the statistical model assumed. Under an FE model the optimal weighting factor is the inverse of the sampling variance of each reliability coefficient, $w_{\mathrm{i}}^{\mathrm{FE}} = 1/V(y_i)$, with $V(y_i)$ being the within-study sampling variance of the reliability coefficient of the i_{th} study. Alternatively, under an FE model the meta-analyst can decide not to weight the reliability estimates, that is, $w_{\mathrm{i}}^{\mathrm{FE}} = 1$. Under an RE

Table 1 Transformation methods for internal consistency coefficients, with back-transformations and sampling variances

	Transformation	Back-transformation	Sampling variance V(y_i)[¶]
No transformation	$\widehat{\alpha}_i$	—	$V\left(\widehat{\alpha}_i\right) = \frac{2J_i\left(1-\widehat{\alpha}_i\right)^2}{(J_i-1)\left\{n_i-2-[(J-2)(k-1)]^{1/4}\right\}}$
Fisher's Z	$Z_i = \frac{1}{2}\ln\left(\frac{1+\widehat{\alpha}_i}{1-\widehat{\alpha}_i}\right)$	$\widehat{\alpha}_i = \frac{e^{2Z_i}-1}{e^{2Z_i}+1}$	$V\left(Z_i\right) = \frac{1}{n_i-3}$
Hakstian-Whalen	$T_i = \sqrt[3]{1-\widehat{\alpha}_i}$	$\widehat{\alpha}_i = 1 - T_i^3$	$V\left(T_i\right) = \frac{18J_i(n_i-1)(1-\widehat{\alpha}_i)^{2/3}}{(J_i-1)(9n_i-11)^2}$
Bonett	$L_i = \ln(1-\lvert\widehat{\alpha}_i\rvert)$	$\widehat{\alpha}_i = 1 - e^{L_i}$	$V\left(L_i\right) = \frac{2J_i}{(J_i-1)(n_i-2)}$

$\widehat{\alpha}_i$: alpha coefficient reported in the i_{th} study. n_i: sample size of the i_{th} study. J_i : number of items of the test version used in the *i*th study. *k*: number of alpha coefficients of the RG meta-analysis. [¶]The sampling variance formula for the untransformed internal consistency coefficients is that proposed by Bonett (2002). ln: natural logarithm. Hakstian-Whalen: Hakstian and Whalen's (1976) transformation. Bonett: Bonett's (2002) transformation model, the optimal weights are defined as the inverse of the sum of the sampling variance and the between-studies variance, $w_{\mathrm{i}}^{\mathrm{RE}} = 1/[V(y_i) + \tau^2]$, with τ^2 being an estimate of the between-studies variance (Borenstein et al., 2009; Cooper et al., 2019). Alternatively, an RE model can be applied by weighting the reliability coefficients by its sample size instead of its inverse variance (Schmidt & Hunter, 2015). In addition to FE and RE models, the varying-coefficient (VC) model (also named *fixed-effects*, in plural, *model*) was proposed in the meta-analytic arena by Laird and Mosteller (1990) and advocated by Bonett (2010) to be applied in RG meta-analysis (see also Bender et al., 2018; Rice et al., 2018). Like the RE model, VC assumes that each individual reliability coefficient is estimating a different population parameter, but contrary to the RE model, VC does not assume that the parametric reliability coefficients are a representative sample of a larger population of potential reliability coefficients. As a consequence, like the FE model, results from the VC model can only be generalized to a set of studies with identical characteristics to those of the studies included in the meta-analysis, and the optimal estimate of the average reliability coefficient implies not weighting the individual coefficients ($w_{\mathrm{i}}^{\mathrm{VC}} = 1$). Unlike the RE model, the VC model assumes that the parametric effect size estimated by each study has not been selected from a hypothetical superpopulation of parametric effect sizes. The mathematical formulation of the three statistical models can be found in Table S1 in Supplementary file 1: https://bit.ly/7rfx65.

Regardless of the statistical model assumed, in an RG meta-analysis it is usual to calculate an average reliability coefficient, its sampling variance, and a 95% confidence interval to estimate the average population reliability coefficient. Veroniki et al. (2019) identified 15 alternative methods for constructing confidence intervals for the average effect size under an RE model. Out of the numerous methods available for constructing confidence intervals, this study focuses on those commonly used in RG meta-analysis. These selected methods are presented in Table 2. The methods differ in terms of whether they transform the reliability estimates, the statistical model assumed, and how they weight reliability coefficients. As a result, we have considered methods under the FE, RE, and VC statistical models, along with methods based on ordinary least squares (OLS). OLS method consists of applying conventional statistical methods, that is, to calculate an unweighted mean of reliability coefficients, to estimate its sampling variance, and to construct a 95% confidence interval as if the reliability estimates were single data from a sample of participants. Although the OLS method can be thought of as an FE model, here we consider it separately, as many RG meta-analyses have applied OLS methods without declaring the statistical model assumed. Note that in OLS and FE methods the reliability coefficients can be transformed or not to normalize their distribution and stabilize variances (in Table 2 the term 'y_i' interchangeably represents the transformed or untransformed reliability coefficient of the *i*th study). Under the VC model advocated by Bonett (2010), the average of the population reliability coefficients

Table 2 Computational formulas to calculate an average reliability coefficient, its sampling variance, and a 95% confidence interval for different statistical models

Model	Average $\left(\overline{y}\right)$	Variance $\left(V(\overline{y})\right)$	Confidence Interval (CI)
OLS	$\overline{Y}_{OLS} = \frac{\sum_i y_i}{k}$	$V(\overline{Y}_{OLS}) = \frac{S_y^2}{k}$	$CI_{OLS} = \overline{Y}_{OLS} \pm \left\lvert t_{k-1,\alpha/2} \right\rvert \sqrt{V(\overline{Y}_{OLS})}$
FE	$\overline{Y}_{FE} = \frac{\sum_i w_i^{FE} y_i}{\sum_i w_i^{FE}}$	$V(\overline{Y}_{FE}) = \frac{1}{\sum_i w_i^{FE}}$	$CI_{FE} = \overline{Y}_{FE} \pm \left\lvert z_{\alpha/2} \right\rvert \sqrt{V(\overline{Y}_{FE})}$
VC	$\overline{Y}_{VC} = \frac{\sum_i \widehat{\alpha}_i}{k}$	$V(\overline{Y}_{VC}) = \frac{\sum_i V(\widehat{\alpha}_i)}{k^2}$	$CI_{VC} = 1 - exp\left[\ln\left(1 - \overline{Y}_{VC}\right) - b \pm \left\lvert z_{\alpha/2} \right\rvert \sqrt{V(\overline{Y}_{VC})/(1 - \overline{Y}_{VC})^2}\right]$
RE	$\overline{Y}_{RE} = \frac{\sum_i w_i^{RE} y_i}{\sum_i w_i^{RE}}$	$V(\overline{Y}_{RE}) = \frac{1}{\sum_i w_i^{RE}}$	$CI_{RE} = \overline{Y}_{RE} \pm \left\lvert z_{\alpha/2} \right\rvert \sqrt{V(\overline{Y}_{RE})}$
REi	$\overline{Y}_{REi} = \frac{\sum_i w_i^{RE} y_i}{\sum_i w_i^{RE}}$	$V(\overline{Y}_{REi}) = \frac{\sum_i w_i^{RE}(y_i - \overline{Y}_{REi})^2}{(k-1)\sum_i w_i^{RE}}$	$CI_{REi} = \overline{Y}_{REi} \pm \left\lvert t_{k-1,\alpha/2} \right\rvert \sqrt{V(\overline{Y}_{REi})}$
REn	$\overline{Y}_{REn} = \frac{\sum_i n_i \widehat{\alpha}_i}{\sum_i n_i}$	$V(\overline{Y}_{REn}) = \frac{\sum_i n_i(\widehat{\alpha}_i - \overline{Y}_{REn})^2}{k\sum_i n_i}$	$CI_{REn} = \overline{Y}_{REn} \pm \left\lvert z_{\alpha/2} \right\rvert \sqrt{V(\overline{Y}_{REn})}$

OLS Ordinary least squares method; *FE* Fixed-effect model; *VC* Varying-coefficient model; *RE* Random-effects model;. *REi* Random-effects model with the improved method of Hartung and Knapp (2001); *REn* Random-effects model weighting by sample size. y_i=transformed or untransformed reliability coefficient of the *i*th study. $\widehat{\alpha}_i$ = untransformed internal consistency reliability coefficient of the *i*th study. n_i=sample size of the ith study. k=number of studies. S_y^2 = variance of the k transformed or untransformed reliability coefficients. $t_{k-1,\alpha/2}$ = (α/2)×100% percentile of the Student *t*-distribution with *k*-1 degrees of freedom. $z_{\alpha/2}$ = (α/2)×100% percentile of the standard normal distribution. $b = ln\left[\overline{n}/(\overline{n} - 1\right]$, *ln* being the natural logarithm and $\overline{n}$ being the harmonic mean of the sample sizes: $\overline{n} = k/\sum(\frac{1}{n_i})$

is estimated by calculating an unweighted average of the untransformed internal consistency coefficients; however, to construct a 95% confidence interval the average reliability coefficient must be transformed by Bonett's method. Table 2 also shows three methods under an RE model. The standard RE method implies estimating the sampling variance of the average reliability coefficient as the inverse of the sum of the weights (w_i^{RE}) and a standard normal distribution to construct a confidence interval (Konstantopoulos & Hedges, 2019). Following Schmidt and Hunter's (2015) approach, the REn method consists of not transforming the reliability coefficients and weighting them by the sample size of each study. Finally, the REi method is based on an improved method proposed by Hartung and Knapp (2001) to estimate the sampling variance of an average effect size and to assume a Student *t*-distribution with degrees of freedom equal to $k - 1$, k being the number of studies, to construct a confidence interval (Sánchez-Meca & Marín-Martínez, 2008). REi method offers better adjustment to the nominal confidence level than the RE and REn methods, as it takes into account uncertainty in estimation of between-studies variance, τ^2 (Hartung & Knapp, 2001; Rubio-Aparicio et al., 2018; Sánchez-Meca & Marín-Martínez, 2008; Veroniki et al., 2019).

Statistical theory predicts OLS methods as exhibiting the largest confidence widths, as they do not take advantage of cumulating the sample sizes of the primary studies when computing a confidence interval for the average reliability coefficient. They are followed by REi method, as it takes into account two sources of error among the reliability estimates (within- and between-study variability) and uncertainty in estimating the between-studies variance. RE and REn methods will offer narrower confidence widths than REi method, as they do not consider uncertainty in the estimation of the between-study variance. The VC method will present narrower confidence widths than the three RE methods, as it does not aim to estimate an average reliability coefficient from a super-population of potential reliability coefficients, but the average population coefficient of the studies included in the RG meta-analysis. Finally, the FE method will exhibit the narrowest confidence widths, as it assumes that the reliability estimates share a common population reliability coefficient (Sánchez-Meca et al., 2013).

Method

This investigation is a meta-review, a methodological review of the RG meta-analyses conducted in psychology aimed at characterizing how measurement error of psychological tools varies from one test application to the next. This study was not preregistered.

Study selection criteria

To be included in this methodological review, studies needed to fulfil the following selection criteria: (a) to be an RG meta-analysis on one or several psychological tools; (b) to report the complete dataset of the individual reliability estimates extracted from the primary studies; (c) to report at least one dataset of internal consistency reliability coefficients (Cronbach's alpha, omega coefficients, parallel-forms, etc.) with at least five individual reliability coefficients, and (d) studies had to be written in English or Spanish. Above all, to be part of our investigation the dataset had to include at least the internal consistency coefficient and sample size of each individual study.

Search strategy

Electronic searches were carried out in the Scopus and EBSCOhost databases. The Google Scholar search engine was also used to broaden the search. The keywords used were "Reliability Generalization", "Meta-Analysis of Internal Consistence" and "Meta-Analysis of Alpha Coefficients". The temporal range was from 1998 to July 2020. The initial date of the search was established due to the seminal article by Vacha-Haase (1998). The full search strategy followed in each database is available in Supplementary file 2: https://bit.ly/x2djy1.

Figure 1 presents a flow diagram outlining the selection process of studies. The electronic searches yielded 385 references. Additional informal searches produced another 30 references. On discarding duplicated references, a total of 239 references were identified as potentially eligible for this research. From these, 207 references were excluded for not fulfilling some inclusion criteria (e.g., methodological studies which did not focus on internal consistency coefficients, did not present the whole dataset with the individual reliability coefficients, the dataset contained less than 5 reliability coefficients, or the psychological tool had only one item). Therefore, 32 RG meta-analyses were included in this research. The references of the 32 RG meta-analyses selected are openly available in Supplementary file 3: https://bit.ly/rkvce1. As many of these studies included several psychological tests, or one psychological test with different subscales, we were able to obtain 138 datasets comprising scales or subscales contributing 4,350 internal consistency coefficients. Although our purpose was to include any type of internal consistency coefficients, all RG meta-analyses selected for this research used only Cronbach's alpha reliability coefficients.

Data extraction

If one RG meta-analysis reported data from more than one psychological scale or the scale had several subscales, we took these as independent datasets for our statistical analyses. Consequently, the 32 RG meta-analyses selected in this methodological review gave a total of 138 datasets of alpha coefficients on psychological scales and subscales. From each dataset, we extracted the alpha coefficients of the primary studies included in each meta-analysis, number of items of each scale/subscale used, sample size, and the mean and standard deviation of test scores.

Data analysis

Statistical methods shown in Table 2 were applied on each of the 138 datasets of alpha coefficients. To estimate the between-studies variance (τ^2), DerSimonian and Laird's (DL) moments method was applied because it is one of the most widely used, although it is not the best one (cf. Blázquez-Rincón et al., 2023; Boedeker & Henson, 2020; Langan et al., 2017; Sánchez-Meca et al., 2013; Sánchez-Meca & Marín-Martínez, 2008; Veroniki et al., 2016; Viechtbauer, 2005). To evaluate the potential impact of the τ^2 estimator on the outcomes of an RG meta-analysis, the restricted maximum likelihood (REML) estimator was also applied. Sensitivity analyses were conducted using ANOVAs, one for the average alpha coefficient and another for its confidence interval width, to compare the meta-analytic results obtained with DL and REML τ^2 estimator. In order to investigate the influence of the τ^2 estimator on the meta-analytic outcomes, two-way ANOVAs with repeated measures for two factors were applied, using the average alpha coefficient and confidence interval width as dependent variables. The two factors considered were the τ^2 estimator (DL vs. REML) and the transformation method. Four transformation methods of the reliability coefficients were considered (not transformation, Fisher's Z, Hakstian and Whalen's and Bonett's transformations) and six statistical models: OLS, FE, VC, and three RE models (standard RE, REi, and REn models). Although a total of 24 combinations could be applied to obtain an average reliability coefficient, only 13 different methods were compared. This is due to the fact that VC (Bonett, 2010) and REn (Schmidt & Hunter, 2015) models do not admit coefficients to be transformed, therefore these statistical methods were applied for untransformed alpha coefficients only. In addition, note that RE and REi methods apply the same formula to calculate an average reliability coefficient (see Table 2). The difference between RE and REi methods is in how to construct a confidence interval. The 13 methods compared to obtain a combined reliability coefficient can be found in Table S2: https://bit.ly/7rfx65.

In addition, 18 different methods to calculate a confidence interval for the average reliability coefficient were applied (see Table 2). Out of these, 16 methods were obtained by combining the statistical models OLS, FE, RE, and REi with the four transformation methods (not transformed, Fisher's Z, Hakstian and Whalen's, and Bonett's transformations). Two additional methods were based on the VC model for Bonett's transformation and the REn model for untransformed coefficients. The 18 methods compared have been described in Table S3: https://bit.ly/7rfx65.

In addition, Shapiro–Wilk's normality test and skewness and kurtosis indices were applied for each of the 138 datasets and on the three transformed coefficients (Fisher's Z, Hakstian and Whalen's, and Bonett's transformations) as well as on those untransformed. This enabled examination of how much the different transformation methods of the internal consistency coefficients achieved the aim of normalizing coefficient distribution.

Another comparison criterion was the amount of heterogeneity exhibited among the alpha coefficients. With this purpose, Q statistic and I^2 index were calculated for the three transformation methods and for the untransformed alpha coefficients in each of the 138 datasets. When applied to an RG meta-analysis, the I^2 index quantifies the amount of true heterogeneity exhibited by a set of alpha coefficients, that is, the variability exhibited by the alpha coefficient that cannot be explained by sampling error, but which is due to the influence of the composition and variability of the study samples and of how each individual study was conducted (Borenstein et al., 2019).

Another way to assess heterogeneity under a RE model is by constructing a prediction interval. Prediction intervals were calculated for each of the coefficient transformations. In an RG meta-analysis a prediction interval estimates the range of values expected for the population reliability coefficient when a new study with similar characteristics to those included in the meta-analysis is conducted (Borenstein, 2019; Borenstein et al., 2019). Theoretically, prediction intervals and confidence intervals should coincide if no heterogeneity between studies is present; in presence of heterogeneity, prediction intervals tend to be wider than confidence intervals (Higgins et al., 2009).

To compare the 13 alternative methods for calculating the average reliability coefficient and the 18 methods for constructing a confidence interval for the average reliability coefficient, two-way ANOVAs were applied. The model included two factors: the assumed statistical model and the transformation of the coefficients, which had four conditions. For the average coefficient estimate, both factors in the ANOVA had four levels, while for the confidence interval width, the statistical model had six conditions. In the event of statistically significant results for any of the factors, post hoc comparisons were performed using Bonferroni's method.

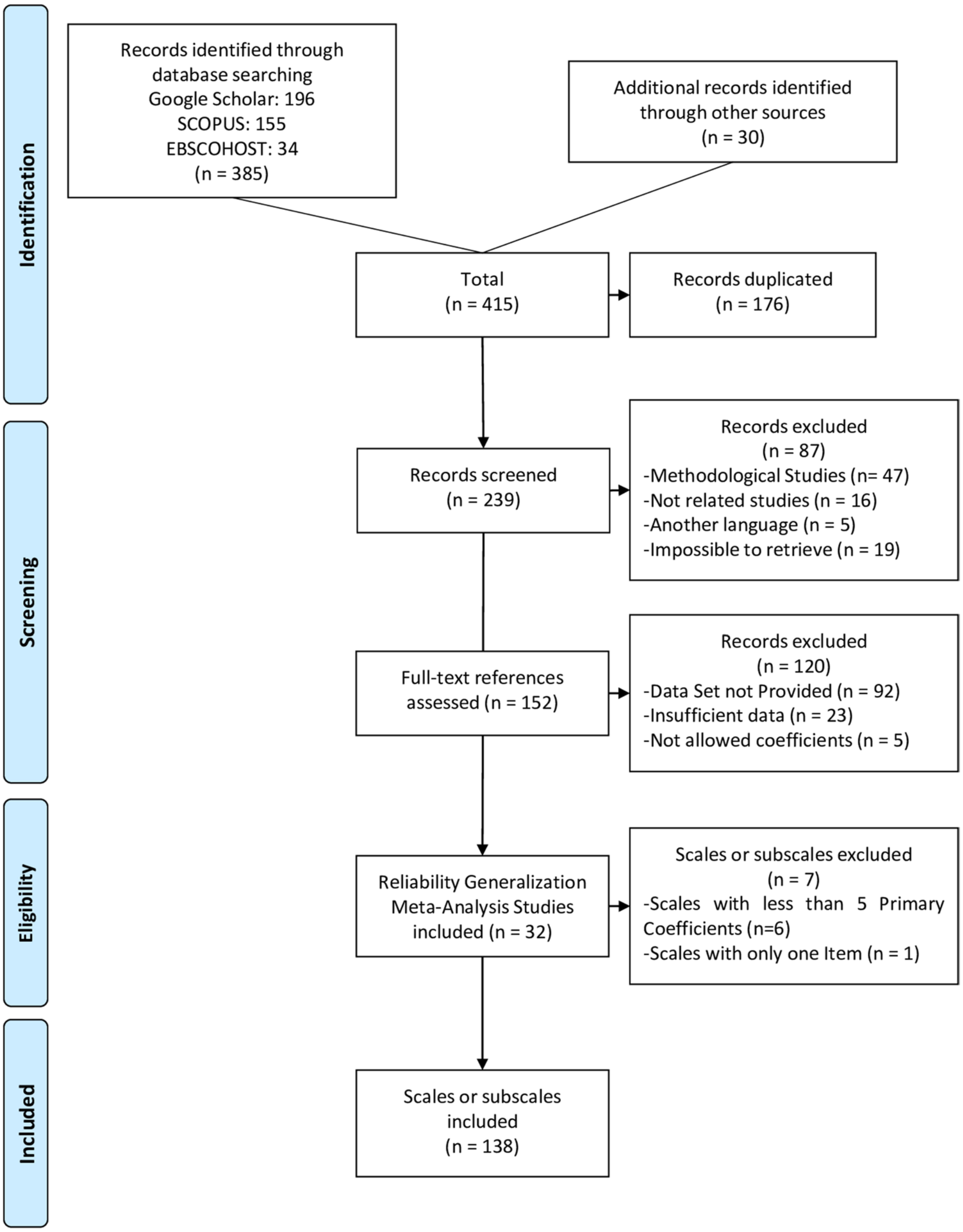
Identification
Screening
Eligibility
Included
Records identified through database searching
Google Scholar: 196
SCOPUS: 155
EBSCOHOST: 34
(n = 385)
Additional records identified through other sources
(n = 30)
Total
(n = 415)
Records duplicated
(n = 176)
Records screened
(n = 239)
Records excluded
(n = 87)
-Methodological Studies (n= 47)
-Not related studies (n = 16)
-Another language (n = 5)
-Impossible to retrieve (n = 19)
Full-text references assessed (n = 152)
Records excluded
(n = 120)
-Data Set not Provided (n = 92)
-Insufficient data (n = 23)
-Not allowed coefficients (n = 5)
Reliability Generalization Meta-Analysis Studies included (n = 32)
Scales or subscales excluded
(n = 7)
-Scales with less than 5 Primary Coefficients (n=6)
-Scales with only one Item (n = 1)
Scales or subscales included
(n = 138)

Fig. 1 Flow diagram of study selection process

The 13 methods for calculating the average reliability coefficient and the 18 methods for constructing a confidence interval for the average reliability coefficient were computed in four different metric scales: the untransformed alpha coefficient and three transformation methods (Fisher's Z, Hakstian and Whalen's, and Bonett's transformations). To ensure comparability, the results for the three transformation methods were back-transformed to the alpha metric using the formulas presented in Table 1.

The 138 meta-analytic datasets as well as the script codes used to analyse them are openly available at: https://bit.ly/vtgf7. All meta-analytic calculations were programmed in R (R Core Team, 2020). Shapiro–Wilk’s normality test and skewness and kurtosis indices were calculated with the R package *moments* (Komsta & Nomovestky, 2015). ANOVAs and post hoc comparisons were carried out with the statistical programs IBM SPSS Statistics (v28; IBM Corp, 2021) and JAMOVI (v.2.2; The Jamovi Project, 2021). Finally, to illustrate the results, multiple violin displays were constructed with the package *ggplot2* in R (Wickham, 2016).

Results

Characteristics of the meta-analytic datasets

The 138 RG datasets were extracted from 32 studies that fulfilled our inclusion criteria. The RG datasets had a number of studies (k) that ranged between 5 and 319 primary studies or alpha coefficients, with an average of 31 primary studies (Median = 14 studies; $Q1 = 9$; $Q3 = 319$). The histogram of the number of studies showed a clear positive asymmetry, with 70.3% of datasets exhibiting fewer than 30 primary studies ($k < 30$) and only 6 datasets (4.3%) with k larger than 100. The distribution of sample sizes for the more than 4,500 alpha coefficients ranged from 38 to 799, with a mean of 209 (Median = 220; $Q1 = 125$; $Q3 = 249$). A summary of the descriptive statistics for both number of studies and sample sizes can be found in Table S4. Both Figure S1 and Table S4 can be found in Supplementary file 1: https://bit.ly/7rfx65.

To transform or not to transform reliability coefficients

One controversial point in the RG meta-analytic arena is whether alpha coefficients should be transformed to normalize their distribution. To examine the extent to which different transformation methods achieved their objective of normalizing the alpha coefficient distribution, Shapiro–Wilk's test and skewness and kurtosis statistics were calculated for each transformation method in each RG dataset. Table 3 presents the results. Regarding untransformed alpha coefficients, almost half of datasets (44.9%) reached statistical significance with Shapiro–Wilk's normality test, indicating a clear departure from the normality assumption. Compared to the untransformed alpha coefficients, the three transformation methods (Fisher's Z, Hakstian and Whalen's, and Bonett's transformations) substantially improved the normality adjustment of the alpha coefficient distribution, with rejection percentages of about 26%. In addition, the skewness indices for untransformed alpha coefficients (Table 3) clearly departed from symmetry (Mean = -0.75; Median = -0.71), whereas transformed coefficients improved the symmetry (Fisher's Z: Mean = 0.005, Media = 0.07; Hakstian and Whalen: Mean = 0.20, Median = 0.14; Bonett: Mean = -0.09, Median = -0.12). To determine whether these differences were statistically significant, a repeated-measures ANOVA was performed. The results confirmed these differences, $F(3, 411) = 31.1$, $p < 0.001$, $\eta^2 = 0.185$. Post hoc comparisons showed differences between no transformation of the coefficients and the three transformations, and between Hakstian and Whalen's and Bonett's transformation. Table S5 in Supplementary File 1 (https://bit.ly/7rfx65) presents the post hoc comparisons.

However, kurtosis indices for untransformed alpha coefficients were close to normality (Mean = 3.74, Median = 2.95), whereas those of the transformed coefficients led to slightly platykurtic distributions (Fisher's Z: Mean = 2.98, Median = 2.61; Hakstian and Whalen: Mean = 3.01, Median = 2.61; Bonett: Mean = 2.94, Median = 2.59). A repeated-measures ANOVA performed to compare the four transformation conditions yielded statistically significant differences, $F(3, 411) = 27.8$, $p < 0.001$, $\eta^2 = 0.169$, specifically between the coefficients without transforming and applying the three transformations. Tables S5 and S6 in Supplementary File 1 (https://bit.ly/7rfx65) presents these results.

Between-study variance estimator

In order to examine whether the choice of the τ^2 estimator in an RE model could affect the average alpha coefficient and the confidence width, a sensitivity analysis was conducted consisting of applying two τ^2 estimators: DL and REML. This comparison only affected to the RE model for the average alpha coefficient and for the RE and REi models for the confidence width and the four transformation methods. The results can be found in Tables S7-S13 in Supplementary File 1 (https://bit.ly/7rfx65). Regarding the average alpha coefficient, using DL or REML τ^2 estimators did not affect the results (see Table S8), $F(1, 137) = 1.11$, $p = 0.294$, $\eta^2 = 0.008$. However, an interaction between the τ^2 estimator and transformation method was found, $F(3, 411) = 26.29$, $p < 0.001$, $\eta^2 = 0.161$. Post hoc comparisons revealed statistically significant differences between the average alpha coefficient for DL and REML τ^2 estimators when alpha coefficients were not transformed (see Table S9). Regarding the confidence width, Table S10 presents the results as a function of the τ^2 estimator (DL vs. REML), transformation method, and statistical model (RE vs. REi). Table S11 presents the results of a three-way ANOVA. No statistically significant differences were found for the τ^2 estimator, $F(1, 137) = 2.12$, $p = 0.147$, $\eta^2 = 0.015$. Like with average alpha coefficient, a statistically significant interaction was found between the τ^2 estimator and transformation method, $F(3, 411) = 4.50$, $p = 0.004$, $\eta^2 = 0.032$, although with negligible proportion of variance accounted for. In fact, any of the post hoc comparisons for this interaction reached statistical significance (see Table S12). Similar results were found for the interaction between τ^2 estimator and statistical model (see Tables S11 and S13). As τ^2 estimator did not affect the results, meta-analytic calculations were presented using DL estimator only.

Averaging a set of reliability coefficients

A total of 13 different methods were applied to average a set of reliability coefficients. In Table 4 some descriptive statistics of the results are shown when an average alpha coefficient was calculated. Both the mean and median indicated that the average alpha coefficients were slightly larger under an FE model without transforming the coefficients, in comparison with the remaining methods. While the lowest average alpha coefficients were found under the OLS method with raw coefficients, the maximum values were found in all transformations within the FE model, with the untransformed coefficients and Hakstian-Whalen's transformation yielding the highest values. The distribution of the average alpha coefficients is shown in multiple violin and boxplots presented in Fig. 2 as a function of statistical model and transformation method.

To compare methods among them, a two-way ANOVA was applied, with the average alpha coefficients as dependent variable and the statistical model and transformation method as factors. The results showed a statistically significant interaction between the two factors, $F(6, 1781) = 3.233$, $p = 0.004$, $\eta^2 = 0.011$, as well as the statistical model, $F(3, 1781) = 8.614$, $p < 0.001$, $\eta^2 = 0.014$. However, the proportion of variance accounted for by these factors was negligible (1.1% and 1.4%, respectively). Bonferroni's post-hoc comparisons indicated that significant differences were found between the FE model and the rest of the models (see Table S14 in Supplementary File 1). Specifically, significant differences were found between the untransformed average coefficients obtained assuming an FE model and the rest of the models, as well as within the FE model itself using Bonett's and Fisher's Z transformations (Table S15 in Supplementary File 1).

Constructing a confidence interval for the average reliability coefficient

Differences among the 18 methods to construct a confidence interval for the average reliability coefficient were also compared in terms of their confidence width. Table 5 presents descriptive statistics obtained by calculating the confidence width across the 18 analytical strategies. Both the mean and median indicated that larger confidence widths were found when OLS method was assumed without transforming the coefficients. While the lowest values were found under an FE model, the maximum values were found under OLS and REi models (i.e., RE model with the improved method of Hartung and Knapp). Figure 3 presents multiple violin and boxplots to illustrate the confidence widths through the different analytic methods compared. A two-way ANOVA was applied on the confidence widths as a function of the statistical model and transformation method. Statistically significant differences were found for the statistical model assumed, $F(5, 2466) = 108.675$, $p < 0.001$, $\eta^2 = 0.181$, but not for the interaction, $F(9, 2466) = 0.347$, $p = 0.959$, $\eta^2 = 0.001$, nor for the transformation method, $F(3, 2466) = 0.532$, $p = 0.66$, $\eta^2 = 0.00$). Regarding the multiple comparisons (see Table S16 in Supplementary File 1), a significant result appears between almost all models. Post hoc comparisons revealed statistically significant differences between all the different statistical models, with three exceptions only: FE vs. VC models, OLS vs. REi, and RE vs. REn.

Assessing heterogeneity

To assess heterogeneity exhibited by a set of alpha coefficients, the I^2 index was calculated for each of the 138 RG datasets and for each transformation method, with the purpose of examining the extent to which different transformation methods lead to different I^2 indices. Table 6 and Fig. 4 show the descriptive statistics of the I^2 indices and their distributions for each of the transformations. On average, I^2 index was over 90% in all transformation methods, except for Fisher's Z (88.21%). There was only one dataset with an I^2 value lower than 25% for Fisher's Z ($I^2 = 14.64\%$). When this I^2 value was deleted from the analyses, the average I^2 for Fisher's Z slightly increased (from 88.21% to 88.75%) and its variability decreased ($SD = 11.50$ and 9.65, respectively). In the remaining datasets and transformation methods all I^2 indices exceeded 25%, and only a few showed I^2 values below 75%, the threshold usually established to assume high heterogeneity. Bonett's and Hakstian and Wallen's transformations performed very similarly. In addition, these two transformation methods yielded I^2 indices with lower variability

Table 3 Shapiro–Wilk's normality test, skewness, and kurtosis for each transformation method of alpha coefficients through the 138 meta-analytic datasets

Transformation method	S-W test	Skewness[¶]		Kurtosis[§]	
	Rejection percentage ($p < 0.05$)	Mean	Median	Mean	Median
No transformation	44.9%	−0.757	−0.736	3.75	2.951
Fisher's Z	26.1%	−0.017	0.066	2.968	2.614
Hakstian-Whalen	26.8%	0.19	0.143	3.007	2.607
Bonett	26.8%	−0.098	−0.12	2.934	2.591

S-W test Shapiro–Wilk's normality test. [¶] Skewness indices equal to 0 indicated perfect symmetry of the distribution. [§] Kurtosis indices equal to 3 indicated adjustment to normality

(Range = 53.01% and 54.63%, respectively) than Fisher's Z and untransformed coefficients (Range = 84.77% and 60.32%, respectively).

To determine whether there were statistical differences in the I^2 indices as a function of the transformation method of the alpha coefficients, a repeated measures ANOVA was performed, finding statistically significant differences, $F(3, 411) = 66.6$, $p < 0.001$, $\eta^2 = 0.327$. Post hoc comparisons revealed statistically significant differences between all the transformation methods, with the exception of Hakstian and Whalen vs. Bonett transformations (see Table S17 in Supplementary File 1). Due to the presence of an outlier I^2 index ($I^2 = 14.64\%$), another repeated measures ANOVA was also performed without it. However, deleting this outlier did not change the ANOVA results.

Heterogeneity was also assessed by calculating 95% prediction intervals. Table 7 presents descriptive statistics for the width of these prediction intervals as a function of the transformation method of the alpha coefficients. As expected, prediction intervals were wider than the confidence intervals (compare Table 5 and 7). Figure 5 shows the distribution of prediction interval widths according to the transformation of the alpha coefficients.

To assess whether the transformation method of the alpha coefficients affected the width of prediction intervals, a repeated measures ANOVA was applied. The results showed statistically significant differences, $F(3, 411) = 43.2$, $p < 0.001$, $\eta^2 = 0.24$. Table S18 in Supplementary file 1 shows the results of post-hoc comparisons, with statistically significant differences between all transformation methods. Larger interval widths were found with Bonett's transformation followed by Fisher's Z and Hakstian and Whalen's transformation.

Discussion

With the purpose of determining the extent to which different statistical methods used to integrate a set of reliability coefficients lead to different results, 138 datasets from 32 RG meta-analyses on psychological tests were analysed by applying multiple statistical methods developed in the meta-analytic arena. Regarding the different transformation methods of the reliability coefficients, our findings revealed that Fisher's Z, Hakstian and Whalen', and Bonett's transformations improved the normality adjustment of coefficient distribution than untransformed coefficients. Although the three transformation methods performed similarly, there are conceptual reasons for not using Fisher's Z to transform internal consistency coefficients like alpha and similar coefficients, as Fisher's Z was devised to transform correlation coefficients, whereas an internal consistency reliability coefficient is not a correlation coefficient, but a squared correlation coefficient (a ratio between true score and total score variance). Fisher's Z is adequate to transform test–retest reliability coefficients or parallel-forms coefficients, as these are calculated as correlation coefficients. For alpha coefficients, Hakstian and Whalen's and Bonett's transformations are most recommendable. Therefore, while there are proponents of not transforming reliability coefficients, it appears that

Table 4 Results of average alpha coefficients for each analytic strategy

		Average Alpha							
Model	Transformation	Mean	*SD*	Min	*Q1*	Median	*Q3*	Max	Range
OLS	No transformation	0.819	0.072	0.595	0.775	0.829	0.873	0.974	0.379
	Fisher's Z	0.832	0.070	0.612	0.786	0.840	0.887	0.986	0.373
	Hakstian-Whalen	0.828	0.070	0.610	0.785	0.837	0.883	0.980	0.369
	Bonett	0.833	0.069	0.618	0.787	0.841	0.888	0.986	0.368
FE	No transformation	0.867	0.063	0.634	0.831	0.865	0.914	1	0.366
	Fisher's Z	0.836	0.074	0.527	0.791	0.839	0.890	0.987	0.460
	Hakstian-Whalen	0.848	0.069	0.621	0.804	0.848	0.901	1	0.378
	Bonett	0.837	0.072	0.544	0.793	0.840	0.891	0.987	0.443
RE/REi	No transformation	0.830	0.070	0.622	0.791	0.836	0.883	0.975	0.353
	Fisher's Z	0.833	0.069	0.620	0.787	0.840	0.887	0.986	0.366
	Hakstian-Whalen	0.832	0.069	0.622	0.789	0.838	0.885	0.980	0.358
	Bonett	0.834	0.068	0.624	0.788	0.842	0.887	0.986	0.361
REn	No transformation	0.826	0.076	0.486	0.783	0.830	0.881	0.974	0.487

OLS Ordinary least squares model; *FE* Fixed-effect model; *RE* Standard random-effects model; *REi* Random-effects model with the improved method of Hartung and Knapp (2001); *REn* Random-effects model weighting by sample size. Results for the VC model were not shown in this Table as they coincide with those of the OLS model with untransformed coefficients. Results for RE and REi models are coincident. *SD* Standard deviation; *Min.* and *Max.* Minimum and Maximum average alpha coefficient; *Q1* and *Q3* Quartiles 1 and 3

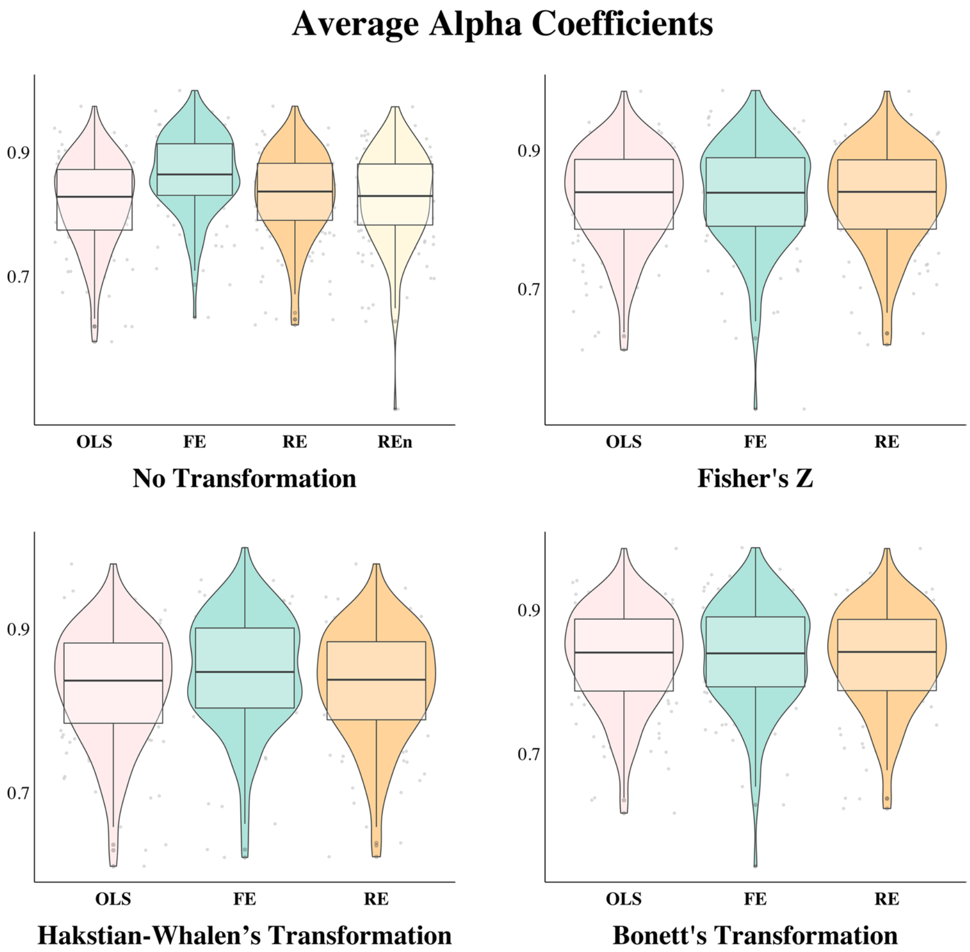

Fig. 2 Multiple violin and box-plots of the 13 different methods for averaging alpha coefficients. *Note:* OLS = Ordinary Least-Squares model. FE = Fixed-Effect model. RE = Standard Random-Effects model weighting by the inverse variance. REn = Random-Effects model weighting by sample size

transformation methods tend to normalize the coefficient distribution. This is recommended since standard meta-analytic methods assume normality in their inferential procedures (cf., e.g., Borenstein et al., 2019; Cooper et al., 2019).

RG meta-analyses always report an average reliability coefficient. Thirteen methods to calculate an average alpha coefficient were compared, depending on the statistical model assumed, weighting factor, and transformation method. An ANOVA applied with the statistical model and transformation of the coefficients as factors showed a statistically significant result for the interaction between them, as well as for the statistical model. However, these factors explained a negligible proportion of the variance (approximately 1%), suggesting a limited influence. Post hoc comparisons indicated that the average alpha coefficients under an FE model were larger than those of other models (Table 4). REn model gave the lowest average alpha coefficients as well as the largest ones (from 0.487 to 0.974), exhibiting the largest variability. REn method consists of weighting the untransformed reliability coefficients by sample size. If reliability coefficients and sample sizes are correlated in an RG meta-analysis, models that include sample size in the weighting factor can yield biased estimates of the average alpha coefficient. Our findings do not allow us to determine the extent to which different statistical models can result in biased estimates of the population alpha coefficient in the presence of alpha-sample size correlation. However, an important recommendation when conducting an RG meta-analysis is to assess the correlation between alphas and sample sizes. If a negative correlation is observed, it may be inferred that this RG meta-analysis may be experiencing what is typically referred to in the field of meta-analysis as 'small study effects,' wherein studies with small sample sizes tend to present higher alpha coefficients than larger ones (Rothstein et al., 2005). Through the 138 RG datasets, correlations between alpha coefficients and sample sizes ranged from -0.79 to 0.73, with Median equal to 0.06 (Mean = 0.07, *SD* = 0.28). These results highlight that it is common to observe both positive and negative correlations between alphas and sample sizes in RG meta-analyses. Therefore, in the presence of such correlations, it is highly advisable to employ techniques for assessing potential biases that can influence the meta-analytic results. Methods like funnel plots, Egger’s test, or trim-and-fill should be utilized to evaluate whether biasing factors, such as 'small study effects,' 'reporting bias,' 'publication bias,' or others, may

Table 5 Results of confidence widths for each analytic strategy

		Confidence width						
Model	Transformation	Mean	*SD*	Min	*Q1*	Median	*Q3*	Max
OLS	No transformation	0.089	0.077	0.015	0.040	0.067	0.105	0.540
	Fisher's Z	0.085	0.077	0.013	0.039	0.059	0.110	0.587
	Hakstian-Whalen	0.085	0.075	0.014	0.038	0.061	0.108	0.572
	Bonett	0.085	0.080	0.013	0.038	0.058	0.110	0.635
FE	No transformation	0.014	0.011	0.000	0.005	0.010	0.020	0.057
	Fisher's Z	0.019	0.015	0.002	0.008	0.013	0.026	0.072
	Hakstian-Whalen	0.015	0.012	0.000	0.006	0.010	0.021	0.060
	Bonett	0.016	0.014	0.001	0.006	0.011	0.022	0.071
RE	No transformation	0.059	0.052	0.009	0.030	0.043	0.072	0.412
	Fisher's Z	0.070	0.057	0.010	0.035	0.054	0.089	0.421
	Hakstian-Whalen	0.068	0.059	0.010	0.034	0.051	0.086	0.460
	Bonett	0.069	0.058	0.010	0.034	0.052	0.089	0.417
REn	No transformation	0.062	0.050	0.010	0.029	0.049	0.081	0.353
REi	No transformation	0.079	0.071	0.011	0.035	0.059	0.099	0.543
	Fisher's Z	0.084	0.077	0.012	0.037	0.058	0.107	0.589
	Hakstian-Whalen	0.082	0.075	0.012	0.036	0.060	0.104	0.573
	Bonett	0.084	0.080	0.012	0.037	0.058	0.107	0.637
VC	Bonett	0.025	0.018	0.002	0.013	0.018	0.032	0.092

OLS Ordinary least squares model; *FE* Fixed-effect model; *RE* Standard random-effects model; *REi* Random-effects model with the improved method of Hartung and Knapp (2001); *REn* Random-effects model weighting by sample size; *VC* Varying-coefficient model; *SD* Standard deviation; *Min.* and *Max.* Minimum and Maximum confidence widths; *Q1* and *Q3* Quartiles 1 and 3

impact the meta-analytic outcomes (Rothstein et al., 2005; Vevea et al., 2019). In cases where a negative correlation exists between alphas and sample sizes, the meta-analyst may choose to apply the FE weighting factor to calculate an average alpha coefficient, as this model is less likely to yield biased estimates.

Conventional RE model weights alpha coefficients by their inverse variance, this being the sum of the sampling variance ($V(y_i)$) and the between-studies variance (τ^2). Note that the between-studies variance is a constant in the RE weighting formula, so that when τ^2 is large in comparison with the sampling variances ($V(y_i)$), the weights become more similar to each other and will therefore approach the OLS method. On the other hand, by including a constant component in the weighting factor will lead to increase the differences between RE and FE models.

A confidence interval for the average reliability coefficient is also typically reported in RG meta-analyses. A total of 18 alternative methods to construct confidence interval for the average alpha coefficient were compared, in terms of the confidence width. Coinciding with previous research, our findings indicated that the different transformation methods of the alpha coefficients barely affected the confidence width for a given statistical model (Romano et al., 2010). However, the statistical model assumed dramatically affected confidence width. ANOVA results showed statistically significant differences as a function of the statistical model, with a proportion of variance accounted for of medium to large magnitude ($\eta^2 = 0.181$).

The largest confidence widths were obtained with the OLS methods, as they do not take advantage of the accumulation of sample sizes through the studies. On average, REi method was that which exhibited confidence widths more similar to those of the OLS method. As expected, the RE and REn methods on average exhibited narrower confidence widths than the REi and OLS methods, with average confidence widths varying between 0.059 and 0.070. Unlike the REi method, the RE and REn methods do not consider the uncertainty in estimating the between-studies variance, providing narrower confidence intervals than those of REi method (Sánchez-Meca & Marín-Martínez, 2008; Sidik & Jonkman, 2002; Stijnen et al., 2021). Both the FE and VC models exhibited the narrowest confidence intervals. The confidence width of VC model was, on average, 0.025, whereas under the FE model the average confidence widths varied between 0.014 and 0.019, being the narrowest widths of all models. The reasons for such narrow confidence widths are different for VC and FE methods. The FE model considers that all studies are estimating a common population reliability coefficient implying that the statistical calculations only take into account one error source: that due to sampling of participants (Borenstein et al., 2009;

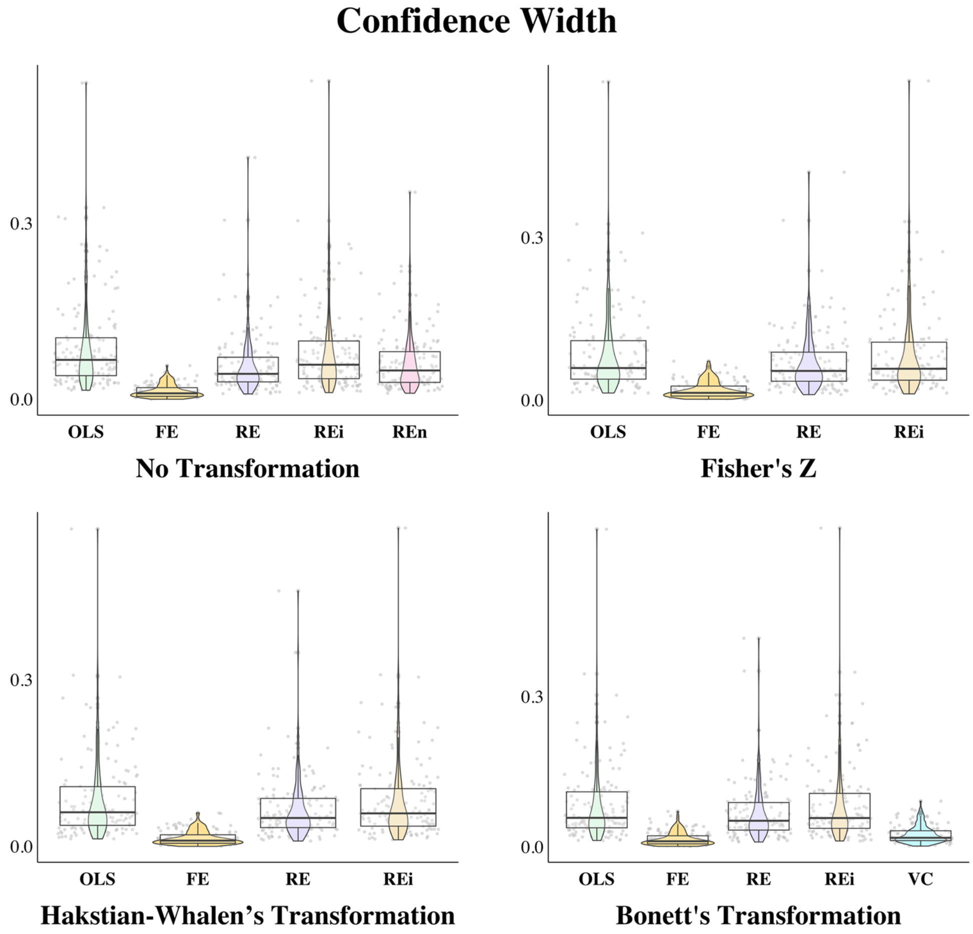

Fig. 3 Multiple violin and boxplots of the 18 different methods for calculating the confidence width. *Note:* OLS: Ordinary Least-Squares model. FE: Fixed-Effect model. RE: Standard Random-Effects model weighting by the inverse variance. REi: Random-Effects model with the improved method of Hartung and Knapp (2001). REn: Random-Effects model weighting by sample size. VC: Varying-Coefficient model

Table 6 Results of aggregating the 138 I^2 indices for each transformation method

	I^2 index						
Transformation method	Mean	*SD*	Min	*Q1*	Median	*Q3*	Max
No transformation	90.833	8.616	39.129	88.845	93.210	96.382	99.452
Fisher's Z	88.212	11.502	14.639	85.672	91.580	95.252	99.413
Hakstian-Whalen	91.700	7.826	45.174	89.741	93.680	96.723	99.797
Bonett	91.693	7.795	46.686	89.652	93.954	96.656	99.698
Fisher's Z¶	88.749	9.653	48.653	85.878	91.583	95.269	99.413

¶Results for Fisher's Z once deleted the dataset with $I^2 = 14.64\%$. *SD* Standard deviation; *Min.* and *Max.* Minimum and Maximum values; *Q1* and *Q3* Quartiles 1 and 3

Sánchez-Meca et al., 2013). The VC model obtains narrower confidence widths than OLS and RE models as this model does not assume that the reliability coefficients from the studies are one random sample of a larger super-population of potential reliability coefficients (Bonett, 2010).

Parameters under the RE model can be estimated using various alternative estimators. Notably, a multitude of between-study variance estimators have been proposed (cf., e.g., Blázquez-Rincón et al., 2023). However, comparing the results of different variance estimators was outside the scope of this study. Nonetheless, we did compare the results of two commonly used between-study variance estimators, the DL and REML estimators. We found negligible differences in the calculation of the average alpha coefficient.

Regarding variability of reliability coefficients, I^2 indices revealed large heterogeneity in most RG datasets, indicating that reliability estimates reported in primary studies are affected by such study characteristics as composition and variability of samples and methods and context of application. In addition, heterogeneity was maintained regardless of the transformation method of the alpha coefficients. Therefore, the search for study characteristics that can explain

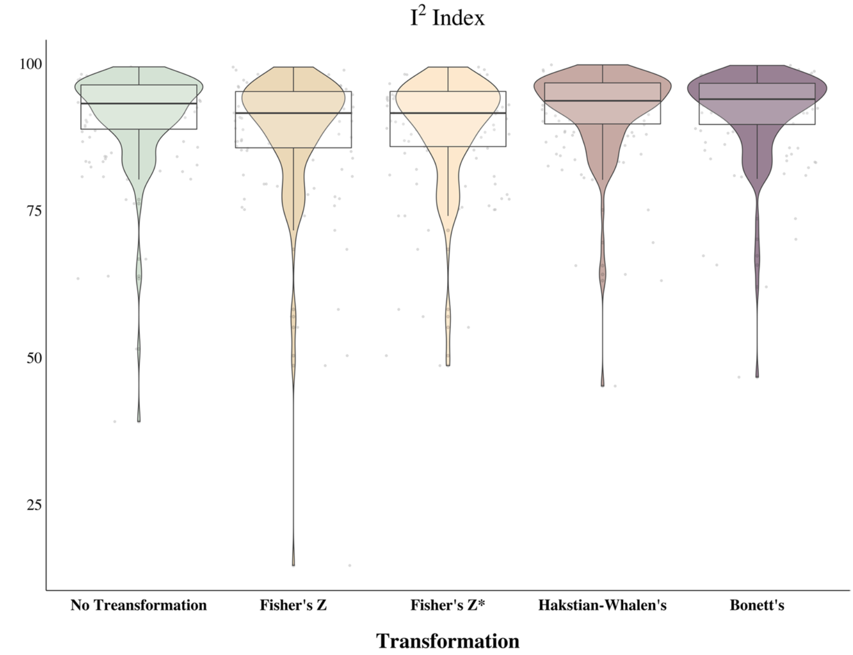

Fig. 4 Multiple boxplots of the I^2 indices for each transformation method. Z* = Fisher's Z once deleted the dataset with $I^2 = 14.64\%$

Table 7 Results of aggregating the 138 prediction intervals width for each transformation method

	95% Prediction interval width						
Transformation method	Mean	*SD*	Min	*Q1*	Median	*Q3*	Max
No transformation	0.207	0.127	0.057	0.127	0.172	0.247	0.998
Fisher's Z	0.273	0.163	0.055	0.168	0.228	0.326	1.037
Hakstian-Whalen	0.254	0.154	0.063	0.160	0.221	0.313	1.205
Bonett	0.297	0.209	0.067	0.177	0.233	0.345	1.333

Hakstian-Whalen: Hakstian and Whalen's transformation. Bonett: Bonett's transformation. *SD* Standard deviation; *Min.* and *Max.* Minimum and Maximum widths. *Q1* and *Q3* Quartiles 1 and 3

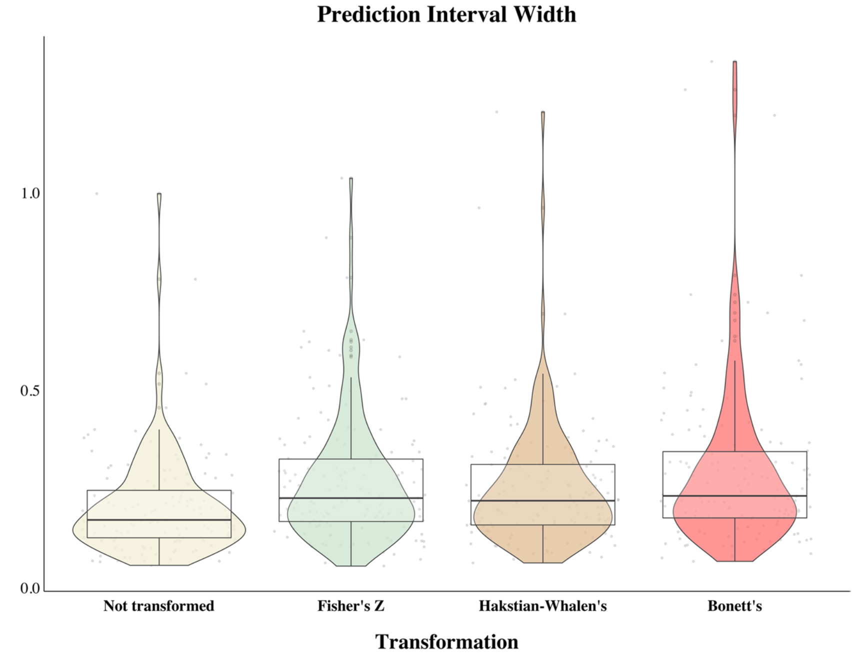

Fig. 5 Multiple boxplots of the widths of the prediction intervals for each transformation method around the 138 RG datasets

heterogeneity is warranted in practically any RG meta-analysis. An additional finding was found about prediction intervals under an RE model. The width of the prediction intervals clearly varied as a function of the transformation method of the alpha coefficients, with wider intervals when Bonett's transformation was applied, followed by Fisher's Z and Hakstian and Whalen's transformation. Therefore, the choice of the transformation method is an important decision to interpret the width of the prediction intervals in an RE model.

How to select the statistical model?

If, as our findings evidence, the selection of the statistical model greatly affects the meta-analytic results, then an important question concerns the arguments that must guide the selection of the statistical model. It is important to note that our investigation does not enable determining which statistical model is most appropriate in an RG meta-analysis, as we have not conducted simulation studies, but empirical research based on real RG datasets. Therefore, our recommendations in this section are not based on our findings, but on previous theoretical work and results of simulation studies. The main question which must guide selection of statistical methods in an RG meta-analysis is to what extent the meta-analyst intends to generalize their results as well as the heterogeneity exhibited by the reliability coefficients. If the aim is to generalize to a set of studies with identical characteristics to those of studies in the meta-analysis, then the FE or the VC models are most recommendable. To decide between FE and VC models, the key question is whether the reliability estimates obtained in the primary studies exhibit heterogeneity. If this is not the case, then the FE model is most appropriate. However, if the reliability estimates exhibit heterogeneity among them, then VC should be chosen. How can we determine whether a set of reliability coefficients are heterogeneous? Several methods can be applied, such as the calculation of the I^2 index, such that if I^2 is larger than 25%, there is evidence of heterogeneity. Another method consists of testing the homogeneity hypothesis with Cochran's Q statistic, such that if the Q statistic reaches statistical significance (e.g., $p < 0.05$) there is evidence of heterogeneity. Other related methods involve calculating a prediction interval around the average reliability coefficient, or interpreting the magnitude of the between-studies standard deviation, τ (Borenstein, 2019; Stijnen et al., 2021). Our results evidenced that RG meta-analyses exhibit large heterogeneity (I^2 indices clearly over 25% and prediction intervals were wider than confidence intervals). As a consequence, FE models will be warranted in exceptional cases only. Even in the presence of apparent homogeneity, applying this model can be risky, as heterogeneity statistics may have limited power when the number of studies is small. Regarding OLS methods, we included it in our comparisons because they have been applied in many RG meta-analyses published in psychology. However, their application in RG meta-analysis, as in other types of meta-analysis, is not recommended under any circumstances. This is because OLS methods do not account for the distributional properties of the reliability coefficients, which can lead to misspecification errors. RG meta-analyses that have estimated their parameters using OLS may yield results that differ significantly from those obtained with RE, VC, or FE models.

When the meta-analyst intends to generalize their results to a larger population of studies with similar but not exactly identical characteristics to those of the studies included in the meta-analysis, then an RE model can be applied. From the three RE models here described, the RE, REi, and REn models, the REi model should be mainly chosen. This is because this model takes into account the uncertainty in estimating the between-studies variance (τ^2). However, to be adequately applied, RE models need several assumptions to be fulfilled: normality of the true reliability coefficient distribution, a stable estimate of the between-studies variance, and random sampling of studies from a larger population of primary studies. Strictly speaking, random sampling assumption cannot be met, as studies included in an RG meta-analysis are never randomly selected from a larger population of potential studies. Nevertheless, it is sufficient if the meta-analyst can reasonably assume, under a conceptual basis, that studies included in an RG meta-analysis are a representative sample of the super-population of primary studies; for example, when there is not correlation between alpha coefficients and sample sizes, or there is not publication bias, small study effects, nor other potential biasing factors (Laird & Mosteller, 1990; Sánchez-Meca et al., 2013). On the other hand, the normality assumption can be relaxed, as recent simulation studies have demonstrated that RE and REi methods are not very affected by departures from normality (Kontopantelis & Reeves, 2012; Rubio-Aparicio et al., 2018). A more serious problem is to obtain an accurate estimate of the between-studies variance (τ^2). A meta-analysis with a small number of studies will have difficulty in accurately estimating τ^2. Note that τ^2 is an important parameter in calculating an average reliability coefficient and to construct confidence intervals and prediction intervals around it. To warrant a stable estimate of τ^2, results from previous simulation studies recommend applying RE and REi methods for meta-analyses with more than 20 studies (Aguinis et al., 2011; Sánchez-Meca et al., 2013). RG meta-analyses with fewer than 20 studies and in the presence of heterogeneity should apply REn method, as it is not necessary to estimate τ^2, provided reliability coefficients and sample sizes are not correlated. Otherwise, the VC model should be the most reasonable choice and the meta-analyst should limit results generalization to studies included in the meta-analysis only.

Finally, it is advisable to apply sensitivity analyses. One of these consists of conducting the statistical analyses both with untransformed and transformed reliability coefficients to assess the strength of findings. In addition, the meta-analyst can apply the leave-one-out technique, consisting of repeating the analyses by deleting one to one each reliability coefficient, with the purpose of identifying outliers. Finally, the correlation between reliability coefficients and sample sizes must always be calculated, as well as constructing a funnel plot, applying Egger's test and, in case of asymmetry of the funnel plot, to apply the trim-and-fill method in order to assess biasing factors related to publication bias and small study effects.

Limitations of study

This investigation has several limitations. Although we were able to analyze a large number of RG datasets (138), they were obtained from 32 RG studies only, a scarce number compared with the approximately 150 RG meta-analyses currently published in psychology. The majority of the RG studies did not report datasets or did not offer the possibility of accessing them. Perhaps due to space limitations in journals, RG meta-analyses with a large number of studies did not report the datasets, such that the RG studies included in our investigation can be a negatively biased sample in terms of number of studies. It is to be expected that, as the transparency and reproducibility principles of the Open Science are implemented in psychological research, meta-analytic databases will be more accessible (Lakens et al., 2016; McNutt, 2014; Pashler & Wagenmakers, 2012). Another limitation was the language, as we only included RG meta-analyses published in English or Spanish. This limitation may impact the generalizability of our results. On the other hand, although we intended to analyze RG datasets of internal consistency coefficients, we were only able to include alpha coefficients. Until now, it has been very rare to find primary studies reporting coefficients other than alpha (e.g., omega, parallel-forms, etc.). However, Cronbach's alpha coefficient has received strong criticism in the last years (Flake & Fried, 2020; Sijtsma, 2009; Yang & Green, 2011), as its very strict assumptions are rarely met in realistic conditions (unidimensionality, tau-equivalence of item factor loadings, uncorrelated errors, multivariate normality). As primary studies report other internal consistency coefficients and other types of reliability (test–retest correlations, inter-rater coefficients), future RG meta-analyses will be able of synthesizing these and then it will be possible to examine the questions considered in this investigation. However, it is reasonable to expect that the majority of our results for alpha coefficients will be applicable to other types of internal consistency coefficients, as well as to other types of reliability, such as temporal stability or inter-rater agreement.

Finally, the main limitation of our investigation is that our findings were derived from empirical comparisons of meta-analytic results using real databases, rather than being based on the results of a simulation study. We designed our study as a preliminary step for conducting future simulation studies that compare the performance of different statistical methods in addressing typical outcomes in an RG meta-analysis. Our results can be valuable for future simulation studies in two ways. First, it was essential to determine whether different analytical methods applied to actual RG meta-analyses exhibit significant differences in meta-analytic results (such as the average reliability coefficient, confidence interval, heterogeneity, etc.). If different statistical methods for synthesizing reliability coefficients show only minor discrepancies, conducting a simulation study may not yield useful insights. Second, our results can assist researchers interested in conducting future simulation studies by allowing them to design manipulated conditions based on the real characteristics of RG meta-analyses typically published in psychology. These characteristics include the number of reliability coefficients, average reliability, sample sizes of individual studies, heterogeneity variance, and more. Consequently, future simulation studies can establish their parameter conditions based on our findings. The descriptive statistics reported in the tables in this paper, as well as in the Supplementary file, will be useful for this purpose.

Future research

The large heterogeneity exhibited in all the RG datasets here analysed evidenced the need to search for study characteristics that can explain at least part of the reliability coefficient variability. Future research should investigate the extent to which different statistical methods to determine the influence of moderator variables reach different results. The statistical methods here compared are based on a univariate approach to RG meta-analysis. Recent methodological work in meta-analysis has developed methods to apply multivariate approaches to RG meta-analyses, such as meta-analytic structural equation modelling (MASEM; Scherer & Teo, 2020). These sophisticated methods require obtaining from each primary study that has applied a given test, the item-item correlation matrix of the test in question, or other statistical data from the factor analyses (factor loadings, residual covariance matrices, etc.). Thus, future research should examine the extent to which univariate and multivariate approaches reach different results when applied to a same RG meta-analysis.

Conclusions

In this research we have demonstrated that the results of an RG meta-analysis are affected conditioned by the statistical model assumed, weighting scheme selected, and other decisions on how to statistically integrate a set of reliability coefficients. Different statistical models estimate different population parameters, so that results are not directly comparable among them. The key point is that the meta-analyst must select the most realistic statistical model, that is, the statistical model that adequately addresses the questions of interest and that better fits the characteristics of the reliability coefficient distribution, their sample composition and variability and sampling framework. Our results also evidence the need for researchers to adhere to the transparency and openness principles of Open Science to guarantee the replicability and reproducibility of psychological research.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12144-023-05604-y.

Authors' contributions Carmen López-Ibáñez (CLI) and Julio Sánchez-Meca (JSM) managed the project, selecting the studies to be analysed and implementing the project. JSM contributed the conceptualisation of the theme and the original idea. JSM also provided the funding. CLI wrote the analysis code and developed all the figures and the visual aspect of the results. Rubén López-Nicolás (RLN) supervised the code script and managed the repository where the files are publicly available. CLI and RLN wrote the original manuscript and analysed and interpreted the results. JSM and Desirée Blázquez-Rincón reviewed the original manuscript and edited it to the final version of the paper. All authors read and approved the final draft of the manuscript.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This research was funded by a grant from the Agencia Estatal de Investigación of the Spanish Government and by FEDER funds (Project nº PID2019-104080GB-I00/MCIN/AEI/10.13039/501100011033).

Data availability The data, materials and analysis script used during the current study are available on the Open Science Framework: https://bit.ly/vtgf7

Declarations

Competing interests The authors have no potential conflict of interest with the material presented in this paper.

References

Aguinis, H., Gottfredson, R. K., & Wright, T. A. (2011). Best-practice recommendations for estimating interaction effects using meta-analysis. *Journal of Organizational Behavior, 32*(8), 1033–1043. https://doi.org/10.1002/job.719

Bender, R., Friede, T., Koch, A., Kuss, O., Schlattmann, P., Schwarzer, G., & Skipka, G. (2018). Methods for evidence synthesis in the case of very few studies. *Research Synthesis Methods, 9*(3), 382–392. https://doi.org/10.1002/jrsm.1297

Blázquez-Rincón, D., Sánchez-Meca, J., Botella, J., & Suero, M. (2023). Heterogeneity estimation in meta-analysis of standardized mean differences when the distribution of random effects departs from normal: A Monte Carlo simulation study. *BMC Medical Research Methodology, 23*(1), 19. https://doi.org/10.1186/s12874-022-01809-0

Boedeker, P., & Henson, R. K. (2020). Evaluation of heterogeneity and heterogeneity interval estimators in random-effects meta-analysis of the standardized mean difference in education and psychology. *Psychological Methods, 25*(3), 346–364. https://doi.org/10.1037/met0000241

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*(4), 335–340. https://doi.org/10.3102/10769986027004335

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods, 15*(4), 368–385. https://doi.org/10.1037/a0020142

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221–237). Russell Sage Foundation.

Borenstein, M. (2019). Heterogeneity in meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 453–468). Russell Sage Foundation.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2019). *Introduction to meta-analysis* (2nd ed.). Wiley.

Botella, J., & Ponte, G. (2011). Effects of the heterogeneity of the variances on reliability generalization: An example with the Beck Depression Inventory. *Psicothema, 23*(3), 516–522.

Botella, J., & Suero, M. (2012). Managing heterogeneity of variances in studies of internal consistency generalization. *Methodology, 8*(2), 71–80. https://doi.org/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*(4), 386–397. https://doi.org/10.1037/a0019626

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2019). *The handbook of research synthesis and meta-analysis* (3rd ed.). Rusell Sage Foundation.

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

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 1–10. https://doi.org/10.1177/2515245920952393

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

Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika, 41*(2), 219–231. https://doi.org/10.1007/BF02291840

Hartung, J., & Knapp, G. (2001). On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine, 20*(12), 1771–1782. https://doi.org/10.1002/sim.791

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*(2), 113–127. https://doi.org/10.1080/07481756.2002.12069054

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (statistics in Society), 172*(1), 137–159. https://doi.org/10.1111/j.1467-985X.2008.00552.x

IBM Corp. (2021). *IBM SPSS Statistics for Windows* (28.0.1.1 (14)) [Windows]. IBM Corp.

Komsta, L., & Nomovestky, F. (2015). *Package 'moments'* [Computer software]. http://www.r-project.org/

Konstantopoulos, S., & Hedges, L. V. (2019). Statistically analyzing effect sizes: Fixed- and random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 245–279). Russell Sage Foundation.

Kontopantelis, E., & Reeves, D. (2012). Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. *Statistical Methods in Medical Research, 21*(4), 409–426. https://doi.org/10.1177/0962280210392008

Laird, N. M., & Mosteller, F. (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care, 6*(1), 5–30. https://doi.org/10.1017/S0266462300008916

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology, 4*(1), 24. https://doi.org/10.1186/s40359-016-0126-3

Langan, D., Higgins, J. P. T., & Simmonds, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: A review of simulation studies: A review of simulation studies. *Research Synthesis Methods, 8*(2), 181–198. https://doi.org/10.1002/jrsm.1198

Mason, C., Allam, R., & Brannick, M. T. (2007). How to meta-analyze coefficient-of-stability estimates: Some recommendations based on Monte Carlo studies. *Educational and Psychological Measurement, 67*(5), 765–783. https://doi.org/10.1177/0013164407301532

McNutt, M. (2014). Reproducibility. *Science, 343*(6168), 229–229. https://doi.org/10.1126/science.1250475

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528–530. https://doi.org/10.1177/1745691612465253

R Core Team. (2020). *R: A language and environment for statistical computing* [Computer software]. https://www.R-project.org/

Rice, K., Higgins, J. P. T., & Lumley, T. (2018). A re-evaluation of fixed effect(s) meta-analysis. *Journal of the Royal Statistical Society Series a: Statistics in Society, 181*(1), 205–227. https://doi.org/10.1111/rssa.12275

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*(3), 306–322. https://doi.org/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*(3), 376–393. https://doi.org/10.1177/0013164409355690

Rothstein, H., Sutton, A., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Wiley.

Rubio-Aparicio, M., López-López, J. A., Sánchez-Meca, J., Marín-Martínez, F., Viechtbauer, W., & Van den Noortgate, W. (2018). Estimation of an overall standardized mean difference in random-effects meta-analysis if the distribution of random effects departs from normal. *Research Synthesis Methods, 9*(3), 489–503. https://doi.org/10.1002/jrsm.1312

Sánchez-Meca, J., Marín-Martínez, F., Núñez-Núñez, R. M., Rubio-Aparicio, M., López-López, J. A., & López-García, J. J. (2019). *Reporting practices in reliability generalization meta-analyses: Assessment with the REGEMA checklist*. XVI Congress of Methodology of the Social and Health Sciences, Madrid, Spain.

Sánchez-Meca, J., López-López, J. A., & López-Pina, J. A. (2013). Some recommended statistical analytic practices when reliability generalization studies are conducted. *British Journal of Mathematical and Statistical Psychology, 66*(3), 402–425. https://doi.org/10.1111/j.2044-8317.2012.02057.x

Sánchez-Meca, J., & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods, 13*(1), 31–48. https://doi.org/10.1037/1082-989X.13.1.31

Sánchez-Meca, J., Marín-Martínez, F., López-López, J. A., Núñez-Núñez, R. M., Rubio-Aparicio, M., López-García, J. J., López-Pina, J. A., Blázquez-Rincón, D. M., López-Ibáñez, C., & López-Nicolás, R. (2021). Improving the reporting quality of reliability generalization meta-analyses: The REGEMA checklist. *Research Synthesis Methods, 12*(4), 516–536. https://doi.org/10.1002/jrsm.1487

Scherer, R., & Teo, T. (2020). A tutorial on the meta-analytic structural equation modeling of reliability coefficients. *Psychological Methods, 25*(6), 747–775. https://doi.org/10.1037/met0000261

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research synthesis* (3rd ed.). Sage.

Sidik, K., & Jonkman, J. N. (2002). A simple confidence interval for meta-analysis. *Statistics in Medicine, 21*(21), 3153–3159. https://doi.org/10.1002/sim.1262

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Stijnen, T., White, I. R., & Schmid, C. H. (2021). Analysis of univariate study-level summary data using normal models. In C. H. Schmid, T. Stijnen, & I. R. White (Eds.), *Handbook of meta-analysis* (pp. 41–64). CRC Press.

The Jamovi Project. (2021). *Jamovi* (2.2) [Computer software]. https://www.jamovi.org

Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues*. Sage.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications*. Sage.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*(1), 6–20. https://doi.org/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*(4), 562–569. https://doi.org/10.1177/0013164402062004002

Veroniki, A. A., Jackson, D., Bender, R., Kuss, O., Langan, D., Higgins, J. P. T., Knapp, G., & Salanti, G. (2019). Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. *Research Synthesis Methods, 10*(1), 23–43. https://doi.org/10.1002/jrsm.1319

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods, 7*(1), 55–79. https://doi.org/10.1002/jrsm.1164

Vevea, J. L., Coburn, C., & Sutton, A. (2019). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 383–429). Russell Sage Foundation.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics, 30*(3), 261–293. https://doi.org/10.3102/10769986030003261

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* [Computer software]. Springer-Verlag. https://ggplot2.tidyverse.org

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*(4), 377–392. https://doi.org/10.1177/0734282911406668

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.