Improving the Reporting Quality of Reliability Generalization Meta-analyses: The REGEMA Checklist

Julio Sánchez-Meca[1], Fulgencio Marín-Martínez[1], José Antonio López-López[1], Rosa Mª Núñez-Núñez[2], María Rubio-Aparicio[3], Juan José López-García[1], José Antonio López-Pina[1], Desirée Mª Blázquez-Rincón[1], Carmen López-Ibáñez[1], and Rubén López-Nicolás[1]

[1] Dept. Basic Psychology & Methodology, University of Murcia (Spain)

[2] Dept. Behavioral & Health Sciences, Miguel Hernández University of Elche (Spain)

[3] Dept. Health Psychology, University of Alicante (Spain)

RUNNING HEAD: The REGEMA checklist

Corresponding author:

Julio Sánchez-Meca

Dept. Basic Psychology & Methodology

University of Murcia (Spain)

jsmeca@um.es

www.um.es/metaanalysis

AUTHOR NOTE:

Portions of this article were presented at the Research Synthesis Conference held in Dubrovnik, Croatia (2019, May). The authors declare that there are no conflicts of interest. Correspondence concerning this article should be addressed to Julio Sánchez-Meca, Faculty of Psychology, University of Murcia, 30100-Murcia, Spain. E-mail: jsmeca@um.es.

FUNDING

Funding for this study was provided by the Ministerio de Economía y Competitividad of the Spanish Government and by funds from the Fondo Europeo de Desarrollo Regional (FEDER; Project No. PSI2016- 77676-P). The funding sponsor had no role in the study design, collection, analysis or interpretation of the data, writing the manuscript, or the decision to submit the article for publication.

ACKNOWLEDGEMENTS

We are very grateful to Raimundo Aguayo (University of Granada, Spain), Juan Botella (Autonomous University of Madrid, Spain), Michael T. Brannick (University of South Florida), John C. Caruso (University of Montana), Salvador Chacón-Moscoso (University of Seville, Spain), Timo Gnambs (Osnabrück University, Germany), Alejandro Guillén-Riquelme (University of Granada, Spain), Chan M. Hellman (University of Oklahoma), Susana Sanduvete-Chaves (University of Seville, Spain), Wim van den Noortgate (University of Louvain, Belgium),Wolfgang Viechtbauer (Maastricht University, Netherlands), and one anonymous reviewer for answering our survey on REGEMA and their useful comments and suggestions.

Improving the Reporting Quality of Reliability Generalization Meta-analyses: The REGEMA Checklist

Abstract

Reliability generalization (RG) is a meta-analytic approach that aims to characterize how reliability estimates from the same test vary across different applications of the instrument. With this purpose RG meta-analyses typically focus on a particular test and intend to obtain an overall reliability of test scores and to investigate how the composition and variability of the samples affect reliability. Although several guidelines have been proposed in the meta-analytic literature to help authors improve the reporting quality of meta-analyses, none of them were devised for RG meta-analyses. The purpose of this investigation was to develop REGEMA (REliability GEneralization Meta-Analysis), a 30-item checklist (plus a flow chart) adapted to the specific issues that the reporting of an RG meta-analysis must take into account. Based on previous checklists and guidelines proposed in the meta-analytic arena, a first version was elaborated by applying the nominal group methodology. The resulting instrument was submitted to a list of independent meta-analysis experts and, after discussion, the final version of the REGEMA checklist was reached. In a pilot study, four pairs of coders applied REGEMA to a random sample of 40 RG meta-analyses in Psychology, and results showed satisfactory inter-coder reliability. REGEMA can be used by: (a) meta-analysts conducting or reporting an RG meta-analysis and aiming to improve its reporting quality; (b) consumers of RG meta-analyses who want to make informed critical appraisals of their reporting quality, and (c) reviewers and editors of journals who are considering submissions where an RG meta-analysis was reported for potential publication.

Keywords: meta-analysis; reliability generalization; reporting quality; systematic reviews; reliability coefficient

Improving the Reporting Quality of Reliability Generalization Meta-analyses: The REGEMA Checklist

1. Introduction

Usually, meta-analyses use such effect size indices as standardized mean differences, mean differences, odds ratios, or correlation coefficients, among others, to investigate many different questions, such as treatment effects, associations between variables, risk or protection factors of a disease, or prevalence rates, among others. [1–3] Reliability generalization is a special kind of meta-analysis that aims to investigate how reliability estimates of test scores vary when the test is applied to different samples. There is an extended idea among researchers and practitioners that reliability is a property inherent to the test and, as a consequence, it remains constant regardless of where the test is applied. However, psychometric theory demonstrates that reliability is not an immutable property of the test, but of the test scores and, therefore, it changes from one application to the next.[4–6] In particular, reliability of test scores can vary as a function of the context of application and the composition and variability of the sample to which it is applied.[7,8] In particular, reliability is known to depend on such characteristics as the length of the test, the variability of the test scores, the target population (e.g., community vs. clinical), or the test version (e.g., original vs. adapted to other languages or cultures). All of these factors, and many others, cause variability across the reliability estimates yielded in different applications of the test. The variability inherent to the reliability of test scores is the reason for many guidelines in Psychology to advise studies to report reliability estimates based on their own data, instead of inducing it from previous applications of the test.[9,10]

1.1 What is a Reliability Generalization Meta-analysis?

If reliability changes from one test application to the next, then meta-analysis is an ideal methodology to investigate factors that affect, or explain, such variability. Vacha-Haase[11] coined the term ‘reliability generalization’ (RG) to refer to this type of meta-analysis. An RG meta-analysis aims: (a) to estimate the average reliability of test scores through the multiple applications of the test to different samples and under different contexts, (b) to assess the extent to which reliability can be generalized from one test application to the next, and (c) if reliability varies across test applications, to identify potential study and sample characteristics than might be statistically associated to the reliability coefficients.[12–14] Thus, an RG meta-analysis is the only kind of meta-analysis where the main effect size indices are reliability coefficients reported in the primary studies that have applied a given test (e.g., internal consistency coefficients, test-retest coefficients, inter-rater correlations, etc.).

A systematic search has revealed more than 150 RG meta-analyses carried out in Psychology between 1998 and 2019.[15] Examples of psychological tests that have been investigated with this methodology are the *Beck Depression Inventory* (BDI),[16] the *Spielberger State-Trait Anxiety Inventory* (STAI),[17] and the *Yale-Brown Obsessive-Compulsive Scale for Children and Adolescents* (CY-BOCS).[18] RG meta-analysis is sometimes applied to specific measures, but also to categories of measures generally, such as the Big Five personality trait measures,[19] or self-report measures of muscle dysmorphia.[20] RG meta-analysis is being applied not only in psychology, but in other social and health sciences, such as in medicine,[21] physical therapy,[22] nursing,[23] sports sciences,[24] education,[25] or marketing.[26] The main purpose of an RG meta-analysis is to inform researchers and practitioners about the expected reliability of the scores from a given test and which test format and application conditions might affect the reliability estimates. Meta-analyses, in general, are being used to inform decision making among researchers, practitioners, and policy makers in disciplines related to the social and health sciences. Another contribution of RG meta-analyses is to inform artifact distributions in other meta-analyses where the meta-analyst intends to correct the effect sizes for such artifacts as measurement error.[27,28] However, the usefulness of meta-analyses is conditioned by their reporting quality. A poorly reported meta-analysis will have limited usefulness for future research and professional practice. This limitation also affects RG meta-analyses. Furthermore, poor reporting quality hampers replicability and goes against the principle of transparency promoted by the Open Science Framework.[29] A systematic review of 150 RG meta-analyses conducted on psychological tests revealed important deficiencies in the reporting quality, especially regarding key methodological aspects in the conduct of an RG meta-analysis.[15]

1.2 Checklists to Improve the Reporting of Meta-analyses

In the last years, guidelines and checklists on how to adequately report meta-analyses have proliferated. Shea, Dubé and Moher[30] identified 23 checklists and three scales of that kind. The most popular checklist in meta-analysis is the PRISMA statement (*Preferred Reporting Items for Systematic reviews and Meta-Analyses*) elaborated by Moher, Liberatti, Tetzlaff, Altman and the PRISMA Group.[31] The PRISMA checklist consists of 27 items to assess the reporting quality of meta-analyses on the effectiveness of interventions. Several adaptations of the original PRISMA checklist have been developed to be applied to other types of meta-analyses, such as PRISMA-IPD for individual participant data meta-analyses,[32] PRISMA-NMA for network meta-analyses,[33] PRISMA-DTA for meta-analyses of diagnostic test accuracy,[34] PRISMA-A for meta-analyses on the effectiveness of acupuncture,[35] PRISMA-Equity for meta-analyses on health equity and social determinants of health.[36] Shea, Grimshaw, Wells et al.[37] developed AMSTAR (*A MeaSurement Tool to Assess systematic Reviews*), a checklist of 11 items to assess the methodological quality of systematic reviews and meta-analyses on the effectiveness of interventions. Instead of reporting quality, AMSTAR focuses on assessing the methodological quality of meta-analyses. Shea, Reeves, Wells et al.[38] elaborated AMSTAR 2, an update of AMSTAR consisting of 16 items. To assess the reporting quality of meta-analyses of observational studies (cohort, case-control, cross-sectional, correlational studies) Stroup, Berlin, Morton et al.[39] developed the MOOSE guideline (*Meta-analysis Of Observational Studies in Epidemiology*), a checklist of 35 items. Recently, Topor et al.[40] have developed NIRO-SR, a checklist for non-interventional systematic reviews and meta-analyses consisting of 68 items. The APA Publications and Communications Board Working Group elaborated MARS (*Meta-Analysis Reporting Standards*), a 74-item guideline aimed at improving the reporting of meta-analyses in psychology.[41] More recently, Appelbaum et al.[9] have updated the MARS guidelines.

In the psychometric arena, Terwee, Mokkink, Knol et al.[42] developed COSMIN (*COnsensus-based Standards for the selection of health status Measurement INstruments*), a wide scope checklist to evaluate the methodological quality of studies on measurement properties consisting of 9 boxes with 5-18 items each. However, this checklist is more suited for primary psychometric studies than for meta-analyses of those. It is also important to note that RG meta-analyses do not only include psychometric studies, but also studies that have applied a given test to a sample of participants with a substantive purpose. In sum, no checklist has been proposed to date that is applicable to appraise the reporting quality of RG meta-analyses.

There is some evidence that the use of guidelines and checklists is improving the reporting quality of systematic reviews and meta-analyses,[43] such that they have been welcomed by the scientific community.

1.3 Peculiarities of RG Meta-analyses

The specific characteristics of RG meta-analyses make the guidelines and checklists above mentioned inappropriate for their application in the RG field. Some of the items in those checklists are irrelevant to RG meta-analyses, and there are important aspects of RG meta-analyses that are not considered in any of the checklists proposed to date. There are also items from existing checklists that are relevant to RG meta-analyses. However, this should not be taken as an argument to deem a checklist specific to RG meta-analyses as unnecessary. In fact, the existing checklists share items among them. The main point to justify developing a specific checklist for RG meta-analyses is their genuine purpose and several methodological peculiarities that need to be considered.

A first peculiarity of an RG meta-analysis refers to the question of interest: to examine how reliability of test scores varies between different applications of the test. Other kinds of meta-analysis focus on estimating the effects of interventions or associations between variables, not on the reliability of test scores.

A second peculiarity relates to the ‘effect sizes’ used in an RG meta-analysis. Whereas other more typical meta-analyses use effect size indices to quantify the effectiveness of interventions or the relationship between variables (e.g., standardized mean differences, odds ratios, correlation coefficients), in an RG meta-analysis the outcomes of interest are the reliability estimates reported in the studies, such as internal consistency, temporal stability or inter-rater agreement. Actually, in a standard RG meta-analysis the statistical integration of the reliability coefficients does not differ from other meta-analyses that integrate such effect sizes as standardized mean differences, risk ratios, or correlations. It is true that transformations are recommended for synthesizing reliability coefficients in order to normalize their distribution and stabilize their variances, but this is also common with other effect size indices (e.g., risk ratios, odds ratios, correlation coefficients, prevalence rates)[1]. A standard RG meta-analysis applies univariate approaches on single reliability estimates extracted from the studies. However, multivariate approaches have also been developed in the field of RG meta-analysis, which entails extracting from each study the item-item correlation or covariance matrix or other statistical information obtained from the factor analyses (e.g., factor loadings, residual variances) conducted to examine the factor structure of the test, and synthesizing them by means of meta-analytic structural equation modeling (MASEM) approaches.[44,45] In these cases, the ‘effect size’ obtained from each study is not a scalar, but the item correlation (or covariance) matrix, or the factor loadings and other results from the factor analysis. These types of ‘effect sizes’ require a

[1] Note that several extensions have been developed to adapt the original PRISMA checklist to other kinds of meta-analysis, such as PRISMA-Equity, PRISMA-Harms, PRISMA-NMA (for network meta-analysis), or PRISMA-A (for acupuncture). These PRISMA extensions were derived to be applied to meta-analyses with specific purposes, although all of them use similar effect size indices (e.g., mean differences, standardized mean differences, risk ratios, odds ratios, etc.) and the statistical integration is similar in all of them. Thus, the similarity of effect sizes and of the statistical integration do not preclude the development of specific checklists for different purposes. A genuine purpose and/or methodological peculiarities can be sufficient reasons for developing a checklist that accommodates that distinctiveness.

special statistical treatment.[46] An example of multivariate RG meta-analysis is that of Scherer, Siddiq, and Tondeur.[47]

Another peculiarity refers to the problem of publication bias. Although publication bias can be due to different causes (e.g., statistical significance, effect magnitude and direction, sample size, etc.), the most common mechanism is linked to the statistical significance, such that it is more likely to publish a paper with $p < .05$ than papers with non-statistically significant results.[48] The consequence of publication bias is an overestimation of the true population effect. As publication bias is mainly a function of the p-values and not of the reliability of the test scores, RG meta-analyses should not be affected by this problem. However, a special type of publication bias typical of RG meta-analyses is the phenomenon of *reliability induction*.[49] This term refers to a very common practice among researchers of inducing reliability from previous applications of the test (e.g., from previous psychometric studies of the test), rather than reporting an estimate based on the scores obtained in the current study. Reliability induction can be considered one of the questionable measurement practices, with the latter defined as "decisions researchers make that raise doubts about the validity of the measures, and ultimately the validity of study conclusions".[50(p456)]

Two main types of reliability induction have been described[51]: reliability induction 'by report' occurs when a study reports some reliability estimate of test scores obtained in another study, whereas reliability induction 'by omission' consists of failing to mention reliability at all even though one or more psychometric instruments were used in the study. In general, the percentage of studies inducing reliability tends to be very large (over 75%). In a systematic review of 100 RG meta-analyses on 123 psychological tests and including

41,824 studies, Sánchez-Meca, Rubio-Aparicio, López-Pina et al.[52] found a total reliability induction rate of 78.6%. As RG meta-analyses can only use studies reporting reliability estimates with the data at hand, a criticism they have received is that they base their conclusions on a subset of the total population of studies that have applied the test. There is also the possibility that reliability induction becomes a special type of reporting bias. If, for example, researchers that obtained low reliability coefficients (e.g., < .7) when applying a given test decided not to report it, then the results of an RG meta-analysis about that test will overestimate the true reliability of the test scores. Therefore, in an RG meta-analysis it is important to assess the potential existence and implications of a biased reporting pattern of the reliability estimates. To cope with this problem, an analytical strategy that we propose is to compare the composition and variability of the samples used in studies that reported genuine reliability estimates with those that induced reliability. It is known that reliability varies as a function of composition and variability of the sample (e.g., target population, mean and standard deviation, SD, of the test scores, gender and ethnic distribution, among others). Therefore, if the reporting and the inducing studies describe samples with similar composition and variability, then it can be reasonably assumed that reliability estimates are similar across both categories and, as a consequence, we can discard reporting bias of reliability as a threat to the validity of the meta-analytic results. If, on the contrary, studies that induced reliability used samples with different composition and variability to those of the studies that reported reliability (e.g., SDs of the test scores systematically lower than those of the studies that reported reliability estimates), then it can be suspected that scores from the inducing studies would have exhibited lower reliability estimates than those from the reporting studies, leading to a problem of reporting bias of the reliability. Therefore, reliability induction can lead to a special type of reporting bias which needs to be addressed by means of analytic strategies not contemplated in other types of meta-analysis.

RG meta-analyses can also be affected by other questionable measurement practices committed by researchers.[50] One of them happens when a researcher modifies the structure of the test (i.e., deleting or adding items) to achieve adequate reliability (e.g., > .7). An RG meta-analysis focused on a test whose structure has been altered in this way in some applications will obtain biased results. Another very frequent questionable practice is to report a reliability coefficient (e.g., an alpha coefficient) without checking whether their assumptions are met. Thus, an RG meta-analysis of alpha coefficients can overestimate or underestimate the true population reliability when their assumptions are not met.[28,46,50,53,54] Another questionable practice refers to a type of selection bias, which can happen when a sample of university students is used in a psychometric study for a test aimed to characterize an attribute on a community or a clinical population. In this case, range restriction of the test scores will lead to a low standard deviation and, as a consequence, a low reliability coefficient. All of these (and other) questionable measurement practices become special types of risk of bias sources that should be assessed in RG meta-analyses.

Finally, RG meta-analyses present some peculiarities regarding the statistical methods applied.[13] Since the reliability coefficient distributions tend to be skewed, some rationale on whether a transformation was applied (and if so, on the choice among the different methods proposed in the literature) should be provided. Additionally, RG meta-analyses have shown a large heterogeneity regarding the weighting scheme of the reliability coefficients and the statistical model assumed. All of these specific aspects must be addressed when reporting an RG meta-analysis. Additionally, multivariate RG meta-analyses apply special statistical approaches based on factor analysis and MASEM methods.[46]

1.4 Purpose

Guidelines for reporting RG meta-analyses should take into account the aforementioned peculiarities of this kind of meta-analysis. To the best of our knowledge, no checklist specifically focused on how to report RG meta-analyses has been proposed to date. Moreover, in a previous study we analyzed the reporting practices of 150 RG meta-analyses in Psychology and we found substantial shortcomings that challenge the replicability of these meta-analyses.[15] As a consequence, the purpose of this investigation was to elaborate a checklist to help researchers report RG meta-analyses. We have named our checklist REGEMA (*REliability GEneralization Meta-Analysis*). The proposed tool consists of 30 items and a flow chart adapted from existing checklists to the characteristics of RG meta-analyses. REGEMA aims to provide a valid tool that meets an existing need in the field of research synthesis, as it is based on guidelines and checklists to report other types of meta-analyses that have recently been published and are widely accepted for routine use nowadays. We also intended to examine the inter-rater agreement and usability of the REGEMA checklist. REGEMA has been conceived to improve the reporting quality of RG meta-analyses, as well as to inform critical appraisal of published RG meta-analyses by readers and to assist reviewers and editors of journals while considering an RG meta-analysis for potential publication. Although the REGEMA checklist has been mainly devised to improve the reporting quality of RG meta-analyses, throughout this paper we include numerous recommendations on how to warrant an acceptable methodological quality when conducting an RG meta-analysis.

2. The REGEMA checklist

To elaborate the REGEMA checklist and its flow chart, the nominal group methodology was applied.[55] In addition, a pilot study of the reliability of REGEMA was conducted to examine inter-rater agreement and potential compliance of existing RG meta-analyses to REGEMA.

2.1 Development of REGEMA

The first step in producing the REGEMA checklist consisted of reviewing all of the guidelines and checklists proposed in the literature to report or conduct meta-analyses. In particular, the following checklists were consulted: PRISMA, MOOSE, AMSTAR, AMSTAR 2, MARS, and the Appelbaum et al.[9] update of MARS. In addition, 80 methodological publications on RG meta-analysis and/or measurement were also consulted.[7,8,11–14,56–61] Supplementary file 1 contains the complete list of methodological references revised. The literature review was accomplished in November and December 2016 and was later updated in December 2018.

A second step involved holding weekly meetings of members of the Meta-analysis Unit (https://www.um.es/metaanalysis/) between January and April 2017. Along these meetings seven members of the research team developed the items that composed the initial version of REGEMA (JSM, FMM, JALP, RMNN, MRA, JJLG, and JALL). The elaboration of each item was discussed until a consensus among the members was reached. The content of each item was conceived to grasp the relevant characteristics in the report of RG meta-analyses. A total of 16, two-hour long meetings were held. As a result, a first version of REGEMA, composed of 30 items, was produced. The checklist was structured in seven dimensions: Title (one item), Abstract (one item), Introduction (two items), Method (15 items), Results (six items), Discussion (four items), and Funding (one item). In addition, a flow chart was produced to illustrate the literature search, screening process, and final selection of studies in an RG meta-analysis. A specific characteristic of the flow chart was the explicit inclusion of data to calculate the total percentage of studies that induced reliability, as well as separate percentages of reliability induction by report and by omission.

A third step consisted of seeking external feedback from meta-analysis experts on the first version of REGEMA. With this aim, a list of 30 researchers specialized in performing RG meta-analyses and/or in the methodology of meta-analysis was elaborated. To select the experts, we identified the most prolific researchers in the RG field as well as the members of the Society for Research Synthesis Methodology (http://www.srsm.org/) with expertise in RG meta-analysis. Then, we contacted the experts via e-mail, sending the initial version of REGEMA and inviting them to make comments, suggestions, and criticisms to the items and the flow chart. Furthermore, for each item we asked them to assess whether it should be maintained or not in the checklist. To facilitate the response process, the REGEMA checklist was sent electronically. Out of the 30 experts, three of them could not be reached due to erroneous e-mail addresses. Two weeks later, a reminder was sent for experts that had not yet responded.

Out of the 27 experts successfully contacted, 12 of them (44.4%) answered our invitation and sent very useful comments and suggestions to improve the quality of REGEMA. Twenty-seven of the 30 items received comments and suggestions related to wording. Twenty of the 30 items were considered for all experts to be relevant for the checklist, six items were considered relevant for 91.7% of the experts consulted, one item was deemed relevant for 83.3% of the experts, and three items received a lower consensus on their relevance: ‘Type of reliability induction’ (58.3%), ‘Data extraction of inducing studies’ (66.7%), and ‘Graphical techniques’ (66.7%). Following discussions held by the REGEMA research team, changes in the wording were introduced for 12 items. Moreover, the item on ‘Graphical techniques’ was deleted from the checklist and its content was included in another item: “Specify the graphical tools used for result display (e.g., flow chart, forest plots, box plots, stem and leaf displays, histograms, scatter plots)”. Although the items ‘Type of reliability induction’ and ‘Data extraction of inducing studies’ did not receive a large consensus from the experts consulted, we decided to maintain them in the checklist, but incorporating the option ‘Not Applicable’ for RG meta-analyses that did not examine reliability induction. As a result from further discussion meetings, an additional item was added to the checklist to incorporate the existence of a protocol prior to the publication of the RG meta-analysis. The inclusion of this item was considered critical to adhere to the transparency, openness, and reproducibility principles of the Open Science Framework.[29,62] The flow chart received minor changes only. The final REGEMA checklist consisted of 30 items.

2.2 Structure of REGEMA

The REGEMA checklist is structured in eight dimensions: Title (one item), Abstract (one item), Introduction (two items), Method (14 items), Results (six items), Discussion (four items), Funding (one item), and Protocol (one item). Table 1 presents the 30 items of REGEMA. Each item contains an explanatory text describing the purpose of that item.

Supplementary file 2 contains a downloadable Word template for researchers to re-use. The form presented in Table 1 includes three potential answers to each item: ‘Yes’ (the RG meta-analysis fulfils that item), ‘No’ (it does not fulfil it), and ‘Unclear’ (there is insufficient evidence to judge on its compliance). Exceptionally, the option ‘Not Applicable’ (NA: the item is not applicable to the meta-analysis under appraisal) was added to Item 9 (‘Estimating the reliability induction’).

INSERT TABLE 1

The REGEMA checklist may be applied with different purposes. It is primarily intended to guide meta-analysts writing an RG meta-analysis to be submitted to a scientific journal. In that case, the response option ‘Unclear’ included in the REGEMA form presented in Table 1 should not be considered. Optionally, when an RG meta-analysis fulfils a given item, then the meta-analyst should include the page (table, appendix, supplementary file, etc.) of the paper that contains the relevant information for that item. The REGEMA checklist could be also applied by readers, reviewers and editors of journals to make critical appraisals of the reporting quality of an RG meta-analysis published or submitted for publication. In those situations, all four response options (‘Yes’, ‘No’, ‘Unclear’, and ‘Not Applicable’) are relevant.

INSERT FIGURE 1

Here we do not describe all 30 items of the checklist, but only those specific to RG meta-analyses. These items justify the need for a new tool in a field with some existing checklists such as PRISMA, MOOSE, or AMSTAR. The remaining items refer to aspects similar to standard meta-analyses and, as a consequence, they do not need additional explanations.

2.2.1 Title

Regarding the title, (Item 1), we recommend to include the term ‘reliability generalization’ such as, for example, in ‘The Maudsley Obsessive-Compulsive Inventory: A reliability generalization meta-analysis’.[63] However, we consider this term as optional. This is because there are other terms that can also help identify this kind of meta-analysis, for example, ‘Reliability of bidimensional acculturation scores: A meta-analysis’.[64] In fact, the term ‘reliability generalization’ is of extended use in Psychology, but not in other related disciplines of the Social and Health Sciences (e.g., ‘The reliability of the Australasian Triage Scale: A meta-analysis’).[65] The REGEMA checklist was devised to be used not only in Psychology, but in the Social and Health Sciences in general.

2.2.2 Introduction

An RG meta-analysis should present in the Introduction section (Item 3: Background) a thorough description of the attribute/s that the test of interest intends to measure, as well as a detailed description of the test/s, including versions and adaptations to other languages and/or cultures. A description of the measurement model that better fits to the structure of the test (e.g., one-factor, multifactor, hierarchical model) should be described. Furthermore, the Introduction section should also clearly state the purposes of the RG meta-analysis (Item 4: Objectives), specifying whether calculating reliability induction rates was an objective. Although estimating the reliability induction rate of a scale is not mandatory in an RG meta-analysis, we recommend examining it to appraise the extent to which the results of an RG meta-analysis can be generalized to all of the studies that have applied the test, regardless of whether they reported or induced reliability. This is an important point, as there is growing evidence[52] that a large proportion of studies using psychometric instruments induce reliability from previous applications, therefore compromising the generalizability of results from RG meta-analyses.

2.2.3 Method

An important peculiarity of RG meta-analyses is that the dependent variables are not effect size indices aimed to quantify a treatment effect or a relationship between variables, but the reliability coefficients reported in the primary studies with the data at hand. This aspect is included in the Item 8 (Reliability reported). The types of reliability and reliability coefficients to be considered in the meta-analysis must be specified. To this respect, meta-analysts must bear in mind that combining measures of different types of reliability (e.g., internal consistency, temporal stability, or inter-coder agreement) is not appropriate, as they assess different types of reliability. Instead, separate meta-analyses should be performed for each type of reliability. Another possibility is to use advanced techniques, such as multivariate meta-analysis, that allow to obtain estimates on different metrics within a single modeling framework.[66]

The most frequently combined index in RG studies is Cronbach’s alpha coefficient, as it is the most commonly reported reliability estimate in primary studies. However, alpha coefficient is an adequate estimator of the population internal consistency only if very strict assumptions are met (unidimensionality of the test, equal factor loadings, uncorrelated residuals, and normality).[67] Further, studies often report indiscriminately alpha coefficients without checking their assumptions. As a consequence, alpha coefficient has received numerous criticisms and other reliability estimates have been proposed in the literature, such as omega coefficients, whose assumptions are more realistic.[54,67–70] Where the test fits to a congeneric one-factor model (in place of the τ-equivalence model), then omega total coefficient is an adequate reliability estimator. Where unidimensionality is not met, then omega hierarchical and other alternative coefficients are more appropriate reliability estimators of the internal consistency. Where the uncorrelated errors assumption is not met, alternative coefficients to omega must be applied.[54,67,71] Therefore, an RG meta-analysis should take into account the measurement model that best fits to the factor structure of the test and select the most appropriate reliability coefficient to carry out the synthesis. Further, meta-analysts should explicitly discuss the important sources of measurement error in a measure or class of measures (e.g., item sampling error versus temporal error) and critically evaluate whether the reliability coefficients commonly reported for a measure adequately capture the important sources of error, instead of taking for granted the researcher's selection (e.g., using indiscriminately the coefficient alpha).[28,70] To accomplish this purpose, meta-analysts should be able to extract from the primary studies the item correlation or covariance matrices or the factor loadings and other indices from the factor analyses. If the studies did not report this information, meta-analysts should request it from the authors.

If an RG meta-analysis intends to examine the extent to which studies that applied a given test induced its reliability from previous applications of the test, then Item 9 (Estimating the reliability induction and other sources of bias) is in order. In this case, it is recommendable to distinguish between reliability induction 'by report' and 'by omission'.[51] Researchers may not report reliability for a variety of reasons, such as oversight or because it was not required by the journal guidelines. But researchers might also choose not to report reliability because the estimate they obtained was very low. In this last scenario, reliability induction is reflecting a problem of reporting bias that can lead to an overestimation of the average reliability coefficient obtained in the RG meta-analysis. Thus, it is advisable to examine the existence of potential reporting biases in RG meta-analyses. To this aim, we suggest extracting study characteristics (e.g., mean and SD of test scores, mean and SD of the age, gender distribution of the sample, target population – community, clinical), not only from the studies that reported reliability, but also from the studies that induced it. This peculiarity of RG meta-analyses is included in Item 10 (Data extraction of inducing studies). Note that this strategy implies a larger effort, as all of the studies that applied the test, regardless of whether they induced or reported reliability estimates, need to be coded. Nonetheless, such additional effort will allow comparisons of sample composition and variability of the studies that reported and induced reliability, in order to determine whether or not the reporting bias scenario is plausible. For instance, Rubio-Aparicio, Núñez-Núñez, Sánchez-Meca et al.'s[72] RG meta-analysis on the *Padua Inventory-Washington State University Revision of Obsessions and Compulsions* (PI-WSUR) stated in the Methods section: " All study characteristics were not only extracted from studies reporting reliability, but also from those that induced it, with the aim of comparing the characteristics of the studies that both reported and induced reliability" (p. 115). It is important to note that studies that induced reliability are not used in the meta-analytic integration of the reliability coefficients. The role we are proposing for studies that induce reliability is to compare their sample characteristics with those of the studies that reported reliability, with the aim of examining whether the meta-analytic results can be generalized to all of the primary studies that have applied the scale. Especially relevant to the investigation of potential reporting biases is the comparison of score SD of reporting and inducing studies, since higher reliability estimates can be expected from samples showing more variability in the test scores.[4]

In addition to the reliability induction as a kind of reporting bias, Item 9 includes other potential sources of bias that an RG meta-analysis should consider. In this vein, the data extraction form should contain some items about potential threats to the validity of the measurement model mentioned above. For example, coefficient alpha or coefficient omega computed from a unidimensional model assume that the item-specific variance is solely measurement error that does not have any impact on external variables. This assumption has recently been met with intense criticism.[28,50,73,74] Thus, an RG meta-analysis might conclude that available reliability coefficients are insufficient to appropriately estimate the overall reliability of measures or to estimate the size of important sources of error.[53,75,76]

The most commonly reported reliability measures are known to present a skewed distribution,[13] such that some researchers recommend transforming them before carrying out the statistical analyses in order to normalize the sampling distribution and/or stabilize sampling variances.[12,77,78] There are also proponents of using raw reliability measures, as transformation may result in biased results, especially in the presence of heterogeneity.[7,57,60,79–81] Item 12 (Transformation method) in the Method section incorporates this information in the REGEMA checklist. For example, Bachner and O'Rourke's[82] RG meta-analysis did not transform reliability coefficients, whereas O'Rourke[83] transformed alpha coefficients into Fisher's Z. Alpha coefficients can also be transformed with Hakstian and Whalen's[84] formula, as in Aguayo, Vargas, de la Fuente, and Lozano.[85] Some RG meta-analyses have also transformed alpha coefficients with

Bonett's[86] formula, such as in Rubio-Aparicio, Badenes-Ribera, Sánchez-Meca et al.[20] As a sensitivity analysis, we recommend analyzing the data both with the transformed and the untransformed reliability coefficients to check whether the results change.

Our experience reviewing RG meta-analyses revealed that many of them did not report the statistical model assumed in the meta-analytic calculations. Item 13 (Statistical model) includes this important information in the Method section. Two statistical models are the most usually applied in meta-analysis: fixed-effect and random-effects models.[13] The fixed-effect model assumes that all studies that have applied a given test obtain reliability estimates of a common parametric reliability coefficient, such that the variability observed among them is due to random sampling error only. Random-effects models assume that the reliability estimate obtained in each study is estimating a different reliability parameter, and that these parameters are a representative sample from a larger super-population of potential reliability parameters. As a consequence, random-effects models take into account two variability sources: variation due to random sampling of participants for each study (also known as sampling variance or within-study variance) and variation due to sampling of studies (known as between-studies variance or heterogeneity variance). A third statistical model not so widely extended is the varying-coefficients model. This model was initially proposed in meta-analysis by Laird and Mosteller[87] and later advocated by Bonett.[57,79] It assumes that the reliability coefficient from each study is estimating a different reliability parameter but, unlike the random-effects model, the varying-coefficients model does not regard the reliability parameters as a random, representative sample from a super-population of reliability parameters.

Recently, more sophisticated models have been proposed in the RG literature, such as using structural equation models in RG meta-analyses.[44–46] A multivariate RG meta-analysis must describe whether item correlation/covariance matrices or factor loadings were synthesized. Several MASEM approaches have been proposed in the literature, such as Raykov and Marcoulides's[88] approach. More recently, in an excellent tutorial Scherer and Teo[46] describe how to apply two-step MASEM and one-step MASEM approaches (TSMASEM and OSMASEM) to item correlation matrices, or a parameter-based MASEM approach. Regardless of the univariate or multivariate meta-analytic approach adopted, specifying the statistical model assumed in an RG meta-analysis is crucial, as the choice of model will determine the statistical analyses to be conducted and their interpretation.

Typical meta-analytic calculations in RG meta-analyses are to obtain an overall reliability estimate, a confidence interval for it, and to analyze the influence of study characteristics by means of subgroup analyses (or ANOVAs) and meta-regression models. Different weighting methods have been proposed in the RG literature to conduct the statistical analyses. Item 14 (Weighting method) includes this information. Use of conventional statistical analyses (i.e., by ordinary least squares estimation) has been suggested in the RG arena before,[7] implying that the reliability coefficients should not be weighted.[82] Another analytical approach is that advocated by Schmidt and Hunter,[27] consisting of assuming a random-effects model and weighting the reliability estimates by sample size.[16] Random-effects models have also been proposed,[12,56] weighting the reliability coefficients by its inverse variance, the latter defined as the sum of the within-study and the between-studies variances.[72] Another weighting scheme is that advocated from the fixed-effect model,[1,2] in this case weighting each reliability coefficient by the inverse of its within-study variance.[89] The weighting scheme should be selected as a function of the statistical model assumed: inverse variance for fixed-effect (with the within-study variance) and random-effects (summing the within-study and the between-studies variances) models and not weighting when the varying-coefficients model is assumed.[13]

Item 17 (Additional analyses) refers to other statistical analyses, such as sensitivity analyses. For example, as there is no consensus on whether reliability coefficients should be transformed or not, a sensitivity analysis could involve conducting separate analyses with the untransformed and transformed reliability coefficients.[18] Another sensitivity analysis aims to examine the existence of reporting bias of the reliability coefficients (see Item 10). In these cases, the composition and variability of the samples from studies that reported and induced reliability is compared.[20]

The statistical software used in the meta-analytic calculations should be reported in the Method section. Item 18 (Software) of the checklist includes this information. This point is especially important in RG meta-analyses because common statistical approaches in meta-analysis apply weighting methods to estimate the parameters of interest and to test for moderators. Moreover, the number of software tools specifically aimed to conduct meta-analytic calculations has increased in recent years, including R packages such as *metafor*[90] or *meta*.[91] Data analyses under the varying-coefficients model can be conducted with the Excel program developed by Krizan.[92] There are also commercial programs such as *Comprehensive Meta-analysis 3.3*[93] or *MetaWin*.[94] To conduct MASEM approaches to RG meta-analysis metaSEM in R is a good option.[95] It is also good practice to provide the scripts created to perform the meta-analytic calculations (e.g., as supplemental material).

2.2.4 Results

The results of the study selection process are included in Item 19. A flow chart describing this process is highly recommended (Figure 1). If the RG meta-analysis intended to estimate the reliability induction rates, this information must be presented in the Results section. Optionally, the characteristics of the studies that induced and reported reliability can be compared, in order to examine the potential existence of differences between these two groups of studies. To this respect, relevant characteristics might include publication year, geographical area of the study and whether the purpose of the study has psychometric or applied main purpose. Comparing studies that induce and report reliability can shed light on the extent to which the meta-analytic results (based on studies that reported reliability only) can be generalized to all of the studies that have applied the test of interest, regardless of whether they reported or not reliability estimates with the data at hand.

Results of the overall reliability coefficient with its confidence/credibility intervals and assessment of the heterogeneity are contemplated in Item 20 (Mean reliability and heterogeneity). Reporting prediction intervals is also recommended to present the expected range of reliability values if a new study applies the test.[96] If reliability coefficients were transformed, then the results should be back-transformed and presented in the original metric of the reliability coefficient to facilitate their interpretation. It is a good idea to use graphical presentations of the results (e.g., forest plots, boxplots, steam-and-leaf plots). If a test is composed by several subscales, the results should be presented separately for the total scale and for each subscale. If an RG meta-analysis has extracted different types of reliability measures (e.g., alpha coefficients, test-retest correlations, inter-rater agreement coefficients), then their results should be presented separately, as it is not appropriate to combine different types of reliability. If the item-item correlation matrices from the primary studies are available, then meta-analytic structural equation models (MASEM) adapted to RG meta-analysis could be applied.[46]

Item 21 ('Moderator analyses') refers to how to present the results of study characteristics and of the samples that can moderate the variability exhibited by the reliability coefficients. Although linear models are usually applied in moderator analyses, it is important to note that some moderator variables have consistent, nonlinear effects on reliability coefficients. A particularly relevant variable here is the analysis of the number of items of the test when an RG meta-analysis includes different versions or adaptations of a test, or different studies have modified the number of items. This is because the number of items is directly involved in the computation of coefficient alpha, and it indirectly affects other internal consistency estimators such as coefficient omega. If the number of items were directly included as a continuous moderator in a meta-regression model, this could lead to inaccurate results if the effects of additional items are not linear. Accordingly, alternative approaches should be used instead. For example, each coefficient alpha could be adjusted downward using the Spearman Brown Prophecy Formula to estimate the reliability of a single item and these 1-item reliabilities meta-analyzed. Other analytical strategies that do not rely on transforming the coefficients include modeling specific numbers of items as a categorical moderator, as an ordinal moderator or using spline meta-regression to account for the nonlinear effects.

Item 23, 'Comparison of inducing and reporting studies', is one of the most specific for RG meta-analyses. As described above (see Item 10, Data extraction of inducing studies), the existence of studies that induced reliability can be hiding a problem of reporting bias. To address this problem, it is advisable to compare the composition and variability of the samples from the studies that induced reliability with those that reported it. It is particularly interesting to compare the SDs of the test scores of the studies that induced and reported reliability. As psychometric theory predicts, the larger the SD of test scores, the larger the reliability estimate.[4] Thus, if studies that induced reliability showed systematically lower SDs of the test scores than those that reported it, then reliability estimates of inducing studies would have probably been lower. This result might be revealing a problem of reporting bias of the reliability. In this case, the overall reliability found in the RG meta-analysis might be overestimating the true overall reliability of the test scores. Other sociodemographic characteristic of the samples can be compared, such as the mean of the test scores, the average age (and its SD), or the gender and ethnic distribution. As an example, Rubio-Aparicio et al.'s[72] RG meta-analysis on the PI-WSUR compared the SDs of inducing and reporting studies. For nonclinical samples, statistically significant differences were found between the average SD of the studies that induced and reported studies, with a lower average SD for inducing studies. This result is compatible with a problem of reporting bias, such that inducing studies might be hiding low reliability coefficients due to low variability of the test scores. In this case, meta-analytic results should be generalized to reporting studies only.

2.2.5 Discussion

In the Discussion section, Items 27 (Implications for clinical practice) and 28 (Implications for future research) refer to the contributions of the RG meta-analysis to the professional practice and the research field. It is advisable to comment the extent to which the scores of the test of interest exhibit good reliability. To this respect, several guidelines have been proposed in the psychometric literature.[97] When assessing internal consistency, Nunnally and Bernstein[98] recommend that reliability coefficients over .80 be considered appropriate for research purposes and over .90 for clinical practice, whereas coefficients over .70 can be considered acceptable for exploratory research. In addition, to assess the clinical relevance of the internal consistency, Cicchetti[99] suggested the following guidelines: unacceptable for coefficients lower than 0.7, fair for the range from 0.7 to 0.8, good for 0.8 to 0.9, and excellent for values over 0.9.

2.2.6 Protocol

Finally, with the aim of adhering to the principles of transparency, openness, and reproducibility of the Open Science Framework, a study protocol should be released before the RG meta-analysis is conducted, using avenues such as PROSPERO (https://www.crd.york.ac.uk/prospero/) or the Open Science Framework (https://osf.io). Item 30 (Protocol) includes this information.

2.2.7 Study selection flow diagram

Meta-analysts should illustrate the search, screening, and selection process of the studies in a flow chart. Figure 1 presents the REGEMA flow chart. It is an adaptation of other flow charts proposed in the meta-analytic literature (e.g., PRISMA flow chart), but adapted to the peculiarities of RG meta-analyses. Supplementary file 3 contains a downloadable Word template of the flow chart for researchers to re-use. Of special interest is the inclusion of data to report the reliability induction rate of the test under evaluation. This flow chart enables the meta-analyst to report the total reliability induction (i.e., the percentage of studies that applied the test and did not report a genuine reliability estimate with regards to the total number of studies that applied the test) and separate percentages of reliability induction by report and by omission. As RG meta-analyses can only integrate studies that reported a reliability estimate with their data at hand, the generalizability of their results can be compromised if there is a large number of studies that induced reliability. Thus, the extent to which reliability induction is a generalized practice regarding a given test is relevant information for RG meta-analyses and for the scientific community as a whole. For example, López-Pina et al.'s[18] meta-analysis on the CY-BOCS found that out the 345 studies that applied this test, only 47 of them reported a reliability coefficient with the data at hand, such that 86.4% of the studies induced the reliability, with 63.8% inducing the reliability 'by omission' and 22.6% inducing it 'by report'.

2.3 Assessing the inter-rater reliability of REGEMA

As mentioned above, REGEMA was devised to help researchers in two ways. On the one hand, it can be applied by researchers to improve the reporting quality of their RG meta-analyses. On the other hand, it can be used by consumers of published RG meta-analyses for critical appraisal of their reporting quality, as well as by reviewers and editors for critical appraisal of RG meta-analyses submitted for publication.

When the REGEMA checklist is used for critical appraisal of reporting quality, its usefulness will depend on the extent to which the assessments exhibit a reasonable reliability in terms of inter-coder agreement. Many of the items of REGEMA are rather complex, as they include several pieces of content. As an example, Item 7 (Data extraction) includes five different types of study characteristics that should be extracted from the studies: (a) sample size/s, mean/s and standard deviation/s of total test scores and subscales (if applicable); (b) sample characteristics (e.g., target population, country, mean age, standard deviation of the age, gender distribution, ethnic distribution, disorder history

–mean and SD in years); (c) test version (e.g., adaptation/version, number of items, reporting format –self-report, clinician); (d) methods (e.g., study design, purpose of the study –psychometric versus applied–, quality checklist); (e) extrinsic characteristics (e.g., publication status, researchers' affiliations, funding source). Achieving good inter-coder reliability with such complex items is very difficult. As a consequence, an empirical assessment of the inter-rater reliability of REGEMA was accomplished. With this purpose, a codebook was produced to guide the decision process on whether an RG meta-analysis fulfils each item. This involved splitting several items into subitems and specifying detailed rules to help decide if a meta-analysis complied with an item. For instance, Item 7 was split into five subitems and compliance with this item was achieved if subitems (a), (b), and (c) were fulfilled. Splitting items into subitems was needed for Items 1 (Title), 2 (Abstract), 3 (Background), 6 (Search strategies), 7 (Data extraction), 19 (Results of the study selection process), 20 (Mean reliability and heterogeneity), and 24 (Data set). Supplementary file 4 contains the codebook produced for applying the REGEMA checklist for critical appraisal of reporting quality. Although all items had three response options ('Yes', 'No', and 'Unclear'; the option 'Not applicable' was added for some items), detailed rules were provided in the codebook to avoid the option 'Unclear'.

To examine inter-coder reliability across different applications of the REGEMA checklist, a sample of 40 published RG meta-analyses was randomly selected from a database of 150 RG meta-analyses of psychological tests identified in a previous systematic review.[15] Supplementary file 6 presents the references of the 40 RG meta-analyses selected. Eight members of our research team (five lecturers and three Ph.D. students, all of them specialized in meta-analysis) served as coders, such that four pairs of coders were randomly formed (JSM/JALL, FMM/RLN, MRA/RMNN, and DMBR/CLI). Next, 10 RG meta-analyses were randomly assigned to each pair of coders. Several meetings were dedicated to familiarize coders with the REGEMA codebook and train them in the use of the REGEMA checklist. Each coder independently applied the REGEMA checklist to each RG meta-analysis and all decisions made by the coders were registered in an electronic database. Disagreements between the coders were resolved by consensus. Percentages of agreement and Cohen's kappa coefficients were calculated for each item of the checklist.

To examine whether the percentages of agreement were similar among the four pairs of coders, logistic regression models were applied, one for each item, taking the agreement (0: disagreement; 1: agreement) between the two coders as the dependent variable and three dummy variables (representing the four pairs of coders) as the predictors. Supplementary file 5 presents the results. With the exception of two items (Items 3 and 7), no statistically significant differences were found in the agreement rates among the four pairs of coders. For Item 3 (Background), percentages of agreement of the four pairs of coders were 90%, 100%, 100%, and 60%, $\chi^2(3) = 10.18$, $p = .017$. For Item 7 (Data extraction), percentages of agreement were 80%, 100%, 100%, and 70%, $\chi^2(3) = 7.92$, $p = .048$. In both cases, the discrepancies were due to a pair of coders that exhibited percentages of agreement lower than those of the remaining three pairs of coders. In fact, when the logistic regressions were repeated deleting that pair of coders, no statistically significant differences were found (Item 3: $\chi^2(2) = 2.27$, $p = .322$; Item 7: $\chi^2(2) = 4.69$, $p = .096$).

Since the agreement rates were found to be generally similar for the four pairs of coders, the percentages of agreement and kappa coefficients were obtained jointly for the

40 RG meta-analyses. Table 2 presents the results of the inter-coder agreement. Percentages of agreement for the 30 items ranged between 80% (Item 5: Selection criteria) and 100% (Item 19: Results of the study selection process; Item 21: Moderator analyses; Item 23: Comparison of inducing and reporting studies; Item 30: Protocol), with a mean agreement of 93%. Kappa coefficients ranged from .28 (Item 8: Reported reliability) to 1 (Items 19, 21, and 23), with a mean of .78. Following Landis and Koch's[100] guidelines, kappa coefficients over .59 were considered satisfactory. Only Items 8 (Reported reliability) and 28 (Implications for future research) did not reach this cut-point, with kappa coefficients of .28 and .47, respectively, besides they exhibited large percentages of agreement of 90% and 87%, respectively. This result was due to the poor performance of kappa when the marginal frequencies are very unbalanced.[101]

INSERT TABLE 2

2.4 Compliance of RG meta-analyses with REGEMA

After resolving inconsistencies between the coders, it was possible to obtain the percentage of compliance of the 40 RG meta-analyses with each item. Table 3 presents the compliance of the RG meta-analyses examined with the REGEMA checklist. Compliance with the checklist items ranged from 0% (Item 30: Protocol) to 95% (Item 4: Objectives; Item 25: Summary of results), with an average compliance of 52%. Less than 50% of the RG meta-analyses met the requirements of 14 items. None of them had published a protocol prior to the final publication (Item 30). Only 2% of the RG meta-analyses extracted data from the studies that induced reliability (Item 10) and compared the characteristics of the studies that induced and reported reliability (Item 23). Only 22% of the meta-analyses described the results of the study selection process, and only 27% of the meta-analyses met the requirements of a good data extraction process (Item 7). Only 30% of the meta-analyses reported additional analyses (Item 17), such as sensitivity analyses (e.g., analysis of reporting bias, results for transformed and untransformed reliability coefficients, leave-one-out analyses). Only 30% of the meta-analyses reported the complete database to warrant the transparency and openness principles of the Open Science Framework (Item 24). Considering percentages of compliance over 90% as excellent, only three items reached this threshold: the types of reliability reported in the Method section (Item 8, 92% of compliance), a clear description of the objectives in the Introduction section (Item 4, 95% of compliance), and a summary of the results in the Discussion section (Item 25, 95% of compliance).

INSERT TABLE 3

3. Concluding remarks

To date, reporting practices of RG meta-analyses have not received enough attention. We developed REGEMA to help authors report this kind of meta-analysis. The REGEMA checklist can be considered as a valid tool to improve the reporting quality of RG meta-analyses, as its elaboration was based on other widely accepted guidelines and checklists proposed in the meta-analytic literature (PRISMA, MOOSE, AMSTAR, AMSTAR 2, MARS), as well as on methodological papers on RG meta-analysis and measurement. The pilot study we have presented on inter-coder agreement yielded reasonably good estimates of inter-coder reliability across 40 applications of the scale. We note that the satisfactory inter-rater agreement found in our pilot study is only an estimate of the inter-rater reliability of the REGEMA checklist, such that these findings are only applicable for this particular set of RG meta-analyses and this particular set of coders. It is also important to note that

REGEMA was devised to improve the reporting quality of RG meta-analyses, although in this paper we have also provided guidelines to improve the methodological quality of RG meta-analyses.

REGEMA can be used by: (a) meta-analysts conducting or reporting an RG meta-analysis, to ensure they will not forget to report any important aspects; (b) consumers of RG meta-analyses, to make critical appraisals on their reporting quality, and (c) reviewers and editors of journals, to help them in the reviewing process on an RG meta-analysis submitted for publication. Thus, journals might want to add the REGEMA checklist to their publication guidelines in order to help authors to report RG meta-analyses. Supplementary files 2 and 3 contain downloadable Word templates of the checklist and the flow chart for researchers to re-use. These templates are also available from the web-site of our research team (www.um.es/metaanalysis), where we will upload any updates of the checklist and new recommendations about how RG meta-analyses must be conducted and reported. Although we have analyzed the inter-rater agreement of the REGEMA checklist and the compliance with REGEMA on a sample of RG meta-analyses in Psychology, this tool was devised to be applied for RG meta-analyses in the Social and Health Sciences in general.

The results of our pilot study have revealed a large variability and, in some cases, poor compliance of published RG meta-analyses with the REGEMA checklist. Our expectation is that adoption of the REGEMA checklist by authors, readers, and reviewers of RG meta-analyses will improve their reporting quality.

HIGHLIGHTS

- Reliability Generalization (RG) meta-analyses intend to explain how measurement error varies from one test application to the next.

- To date, no checklist had been devised to help researchers conducting and reporting an RG meta-analysis.

- We have developed the REGEMA checklist (REliability GEneralization Meta-Analysis) to improve the reporting quality of RG meta-analyses.

- REGEMA is easy to implement and exhibited satisfactory inter-coder agreement.

- REGEMA can also be applied by readers and editors/reviewers to make critical appraisal of RG meta-analyses.

Data Availability Statement:

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

1. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. *Introduction to Meta-Analysis*. Wiley; 2011.

2. Cooper H, Hedges LV, Valentine JC. *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation; 2019.

3. Schmid CH, Stijnen T, White I. *Handbook of Meta-Analysis*. CRC Press; 2020.

4. Crocker LM, Algina J. *Introduction to Classical and Modern Test Theory*. Holt, Rinehart, & Winston; 1986.

5. Gronlund NE, Linn, Linn RL. *Measurement and Assessment in Teaching. Eighth Edition*. 6th ed. Macmillan; 1990.

6. Traub RE. *Reliability for the Social Sciences: Theory and Applications*. SAGE Publications; 1994.

7. Henson RK, Thompson B. Characterizing measurement error in scores across tudies: Some recommendations for conducting "reliability generalization" studies. *Meas Eval Couns Dev*. 2002;35(2):113-127.

8. Vacha-Haase T, Henson RK, Caruso JC. Reliability Generalization: Moving toward Improved Understanding and Use of Score Reliability. *Educ Psychol Meas*. 2002;62(4):562-569. doi:10.1177/0013164402062004002

9. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *Am Psychol*. 2018;73(1):3-25. doi:10.1037/amp0000191

10. Wilkinson L. Statistical methods in psychology journals: Guidelines and explanations. *Am Psychol*. 1999;54(8):594-604. doi:10.1037/0003-066X.54.8.594

11. Vacha-Haase T. Reliability Generalization: Exploring Variance in Measurement Error Affecting Score Reliability Across Studies. *Educ Psychol Meas.* 1998;58:6-20. doi:10.1177/0013164498058001002

12. Rodriguez MC, Maeda Y. Meta-analysis of coefficient alpha. *Psychol Methods*. 2006;11(3):306-322. doi:10.1037/1082-989X.11.3.306

13. Sánchez-Meca J, López-López JA, López-Pina JA. Some recommended statistical analytic practices when reliability generalization studies are conducted. *Br J Math Stat Psychol*. 2013;66(3):402-425. doi:10.1111/j.2044-8317.2012.02057.x

14. Thompson B. *Score Reliability: Contemporary thinking on reliability issues*. Sage; 2003.

15. Sánchez-Meca J, Marín -Martínez F, Núñez-Núñez RM, Rubio-Aparicio M, López-López JA, López-García JJ. Reporting practices in reliability generalization meta-analyses: Assessment with the REGEMA checklist. In: XVI Congress of Methodology of the Social and Health Sciences; 2019 July; Madrid, Spain.

16. Yin P, Fan X. Assessing the Reliability of Beck Depression Inventory Scores: Reliability Generalization across Studies. *Educ Psychol Meas*. 2000;60(2):201-223. doi:10.1177/00131640021970466

17. Barnes LL, Harp D, Jung WS. Reliability generalization of scores on the Spielberger state-trait anxiety inventory. *Educ Psychol Meas*. 2002;62(4):603-618. doi:10.1177/0013164402062004005

18. López-Pina JA, Sánchez-Meca J, López-López JA, et al. Reliability generalization study of the Yale-Brown Obsessive-Compulsive Scale for children and adolescents. *J Pers Assess*. 2015;97(1):42-54. doi:10.1080/00223891.2014.930470

19. Viswesvaran C, Ones DS. Measurement error in “Big Five factors” personality assessment: Reliability generalization across studies and measures. *Educ Psychol Meas*. 2000;60(2):224-235. doi:10.1177/00131640021970475

20. Rubio-Aparicio M, Badenes-Ribera L, Sánchez-Meca J, Fabris MA, Longobardi C. A reliability generalization meta-analysis of self-report measures of muscle dysmorphia. *Clin Psychol Sci Pract*. 2020;27(1). doi:10.1111/cpsp.12303

21. Brannick MT, Erol-Korkma z HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. *Med Educ*. 2011;45(12):1181-1189. doi:10.1111/j.1365-2923.2011.04075.x

22. Meseguer-Henarejos A-B, Sánchez-Meca J, López-Pina J-A, Carles-Hernández R. Inter- and intra-rater reliability of the Modified Ashworth Scale: a systematic review and meta-analysis. *Eur J Phys Rehabil Med*. 2018;54(4):576-590. doi:10.23736/S1973-9087.17.04796-7

23. Barlow KM, Zangaro GA. Meta-analysis of the reliability and validity of the Anticipated Turnover Scale across studies of registered nurses in the United States. *J Nurs Manag*. 2010;18(7):862-873. doi:https://doi.org/10.1111/j.1365-2834.2010.01171.x

24. Martínez-Romero MT, Ayala F, De Ste Croix M, et al. A Meta-Analysis of the Reliability of Four Field-Based Trunk Extension Endurance Tests. *Int J Environ Res Public Health*. 2020;17(9):3088. doi:10.3390/ijerph17093088

25. Yeo S. Reliability Generalization of Curriculum-Based Measurement Reading Aloud: A Meta-Analytic Review. *Exceptionality*. 2011;19(2):75-93. doi:10.1080/09362835.2011.562094

26. Homburg C, Klarmann M, Reimann M, Schilke O. What Drives Key Informant Accuracy? *J Mark Res*. 2012;49(4):594-608. doi:10.1509/jmr.09.0174

27. Schmidt, FL, Hunter JE. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. 3rd ed. Sage; 2015.

28. Wiernik BM, Dahlke JA. Obtaining Unbiased Results in Meta-Analysis: The Importance of Correcting for Statistical Artifacts. *Adv Methods Pract Psychol Sci*. 2020;3(1):94-123. doi:10.1177/2515245919885611

29. Nosek BA, Alter G, Banks GC, et al. Promoting an open research culture: author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science*. 2015;348:1422-1425. doi:10.1126/science.aab2374

30. Shea B, Dubé C, Moher D. Assessing the Quality of Reports of Systematic Reviews: The QUOROM Statement Compared to Other Tools. In: *Systematic Reviews in Health Care*. John Wiley & Sons, Ltd; 2008:122-139. doi:10.1002/9780470693926.ch7

31. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *J Clin Epidemiol*. 2009;62:1006-1012. doi:10.1016/j.jclinepi.2009.06.005

32. Stewart LA, Clarke M, Rovers M, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Individual Participant Data: The PRISMA-IPD Statement. *JAMA*. 2015;313(16):1657. doi:10.1001/jama.2015.3656

33. Hutton B, Salanti G, Caldwell DM, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Ann Intern Med*. 2015;162(11):777-784. doi:10.7326/M14-2385

34. McInnes MD, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *J Am Med Assoc*. 2018;319(4):388-396. doi:10.1001/jama.2017.19163

35. Wang X, Chen Y, Liu Y, et al. Reporting items for systematic reviews and meta-analyses of acupuncture: the PRISMA for acupuncture checklist. *BMC Complement Altern Med*. 2019;19(1):208. doi:10.1186/s12906-019-2624-3

36. Welch V, Petticrew M, Petkovic J, et al. Extending the PRISMA statement to equity-focused systematic reviews (PRISMA-E 2012): explanation and elaboration. *Int J Equity Health*. 2015;14(1):92. doi:10.1186/s12939-015-0219-2

37. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol*. 2007;7(1):10. doi:10.1186/1471-2288-7-10

38. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*. 2017;358:j4008. doi:10.1136/bmj.j4008

39. Stroup DF. Meta-analysis of Observational Studies in Epidemiology: A Proposal for Reporting. *JAMA*. 2000;283(15):2008. doi:10.1001/jama.283.15.2008

40. Topor M, Pickering JS, Barbosa Mendes A, et al. *An Integrative Framework for Planning and Conducting Non-Interventional, Reproducible, and Open Systematic Reviews (NIRO-SR)*. MetaArXiv; 2020. doi:10.31222/osf.io/8gu5z

41. APA Publications & Communications Board Working Group on Journal Article Reporting Standards. Reporting standards for research in psychology: Why do we need them? What might they be? *Am Psychol*. 2008;63(9):839-851. doi:10.1037/0003-066X.63.9.839

42. Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res*. 2012;21(4):651-657. doi:10.1007/s11136-011-9960-1

43. Panic N, Leoncini E, De Belvis G, Ricciardi W, Boccia S. Evaluation of the endorsement of the preferred reporting items for systematic reviews and meta-analysis (PRISMA) statement on the quality of published systematic review and meta-analyses. *PloS One*. 2013;8(12). doi:10.1371/journal.pone.0083138

44. Cheung MW-L, Cheung SF. Random-effects models for meta-analytic structural equation modeling: review, issues, and illustrations. *Res Synth Methods*. 2016;7(2):140-155. doi:https://doi.org/10.1002/jrsm.1166

45. Jak S. *Meta-Analytic Structural Equation Modelling*. Springer; 2015. doi:10.1007/978-3-319-27174-3_1

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

47. Scherer R, Siddiq F, Tondeur J. All the same or different? Revisiting measures of teachers' technology acceptance. *Comput Educ*. 2020;143:103656. doi:10.1016/j.compedu.2019.103656

48. Vevea JL, Coburn K, Sutton AJ. Publication bias. In: *The Handbook of Research Synthesis and Meta-Analysis*. 3rd ed. Russell Sage Foundation; 2019:383-429.

49. Vacha-Haase T, Kogan LR, Thompson B. Sample Compositions and Variabilities in Published Studies versus Those in Test Manuals: Validity of Score Reliability Inductions. *Educ Psychol Meas*. 2000;60(4):509-522. doi:10.1177/00131640021970682

50. Flake JK, Fried EI. Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Adv Methods Pract Psychol Sci*. 2020;3(4):456-465. doi:10.1177/2515245920952393

51. Shields AL, Caruso JC. A Reliability Induction and Reliability Generalization Study of the Cage Questionnaire. *Educ Psychol Meas*. 2004;64(2):254-270. doi:10.1177/0013164403261814

52. Sánchez-Meca J, Rubio-Aparicio M, López-Pina J, Núñez-Núñez RM, Marín-Martínez F. The phenomenon of reliability induction in the social and health sciences. In: XIV Congress of Methodology of the Social and Health Sciences; 2015 July; Palma de Mallorca, Spain.

53. Le H, Schmidt FL, Putka DJ. The Multifaceted Nature of Measurement Artifacts and Its Implications for Estimating Construct-Level Relationships. *Organ Res Methods*. 2009;12(1):165-200. doi:10.1177/1094428107302900

54. Yanyun Yang, Green SB. Coefficient Alpha: A Reliability Coefficient for the 21st Century? *J Psychoeduc Assess*. 2011;29(4):377-392. doi:10.1177/0734282911406668

55. Fink A, Kosecoff J, Chassin M, Brook RH. Consensus methods: characteristics and guidelines for use. *Am J Public Health*. 1984;74(9):979-983. doi:10.2105/ajph.74.9.979

56. Beretvas SN, Pastor DA. Using mixed-effects models in reliability generalization studies. *Educ Psychol Meas*. 2003;63(1):75-95. doi:10.1177/0013164402239318

57. Bonett DG. Varying coefficient meta-analytic methods for alpha reliability. *Psychol Methods*. 2010;15(4):368-385. doi:10.1037/a0020142

58. Botella J, Suero M, Gambara H. Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychol Methods*. 2010;15(4):386-397. doi:10.1037/a0019626

59. Brannick MT, Zhang N. Bayesian meta-analysis of coefficient alpha. *Res Synth Methods*. 2013;4(2):198-207. doi:10.1002/jrsm.1075

60. Thompson B, Vacha-Haase T. Psychometrics is Datametrics: the Test is not Reliable. *Educ Psychol Meas*. 2000;60(2):174-195. doi:10.1177/0013164400602002

61. Vacha-Haase T, Thompson B. Score Reliability: A Retrospective Look Back at 12 Years of Reliability Generalization Studies. *Meas Eval Couns Dev*. 2011;44(3):159-168. doi:10.1177/0748175611409845

62. Lakens D, Hilgard J, Staaks J. On the reproducibility of meta-analyses: six practical recommendations. *BMC Psychol*. 2016;4(1):24. doi:10.1186/s40359-016-0126-3

63. Sánchez-Meca J, López-Pina JA, López-López JA, Marín-Martínez F, Rosa-Alcázar AI, Gómez-Conesa A. The Maudsley Obsessive-Compulsive Inventory: A reliability generalization meta-analysis. *Int J Clin Health Psychol*. 2011;11(3).

64. Huynh Q-L, Howell RT, Benet-Martínez V. Reliability of bidimensional acculturation scores: A meta-analysis. *J Cross-Cult Psychol*. 2009;40(2):256-274. doi:10.1177/0022022108328919

65. Ebrahimi M, Heydari A, Mazlom R, Mirhaghi A. The reliability of the Australasian Triage Scale: a meta-analysis. *World J Emerg Med*. 2015;6(2):94-99. doi:10.5847/wjem.j.1920-8642.2015.02.002

66. López-López JA, Page MJ, Lip sey MW, Higgins JPT. Dealing with effect size multiplicity in systematic reviews and meta-analyses. *Res Synth Methods*. 2018;9(3):336-351. doi:10.1002/jrsm.1310

67. McNeish D. Thanks coefficient alpha, we'll take it from here. *Psychol Methods*. 2018;23(3):412.

68. Raykov T, Marcoulides GA. Scale reliability evaluation under multiple assumption violations. *Struct Equ Model Multidiscip J*. 2016;23(2):302-313.

69. Sijtsma K. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*. 2009;74(1):107-120. doi:10.1007/s11336-008-9101-0

70. Revelle W, Condon DM. Reliability. In: *The Wiley Handbook of Psychometric Testing*. Wiley; 2018:709-749. doi:10.1002/9781118489772.ch23

71. Bentler PM. Covariate-free and Covariate-dependent Reliability. *Psychometrika*. 2016;81(4):907-920. doi:10.1007/s11336-016-9524-y

72. Rubio-Aparicio M, Núñez-Núñez RM, Sánchez-Meca J, López-Pina JA, Marín-Martínez F, López-López JA. The Padua Inventory–Washington State University Revision of Obsessions and Compulsions: A Reliability Generalization Meta-Analysis. *J Pers Assess*. 2020;102(1):113-123. doi:10.1080/00223891.2018.1483378

73. Credé M, Harms PD. 25 years of higher-order confirmatory factor analysis in the organizational sciences: A critical review and development of reporting recommendations. *J Organ Behav*. 2015;36(6):845-872. doi:10.1002/job.2008

74. Mõttus R, Bates T, Condon DM, Mroczek D, Revelle W. Leveraging a more nuanced view of personality: Narrow characteristics predict and explain variance in life outcomes. *PsyArXiv*. 2017. doi:10.31234/osf.io/4q9gv

75. Gnambs T. Facets of measurement error for scores of the Big Five: Three reliability generalizations. *Personal Individ Differ*. 2015;84:84-89. doi:10.1016/j.paid.2014.08.019

76. Viswesvaran C, Ones DS, Schmidt FL. Comparative analysis of the reliability of job performance ratings. *J Appl Psychol*. 1996;81(5):557-574. doi:10.1037/0021-9010.81.5.557

77. Feldt LS, Charter RA. Averaging Internal Consistency Reliability Coefficients. *Educ Psychol Meas*. 2016;66:215-227. doi:10.1177/0013164404273947

78. Sawilowsky SS. Psychometrics versus Datametrics: Comment on Vacha-Haase's "Reliability Generalization" Method and Some Epm Editorial Policies. *Educ Psychol Meas*. 2000;60(2):157-173. doi:10.1177/00131640021970439

79. Bonett DG. Meta-analytic interval estimation for bivariate correlations. *Psychol Methods*. 2008;13(3):173-181. doi:10.1037/a0012868

80. Leach LF, Henson RK, Odom LR, Cagle LS. A Reliability Generalization Study of the Self-Description Questionnaire. *Educ Psychol Meas*. 2006;66(2):285-304. doi:10.1177/0013164405284030

81. Mason C, Allam R, Brannick MT. How to meta-analyze coefficient-of-stability estimates: Some recommendations based on Monte Carlo studies. *Educ Psychol Meas*. 2007;67(5):765-783. doi:10.1177/0013164407301532

82. Bachner YG, O'Rourke N. Reliability generalization of responses by care providers to the Zarit Burden Interview. *Aging Ment Health*. 2007;11(6):678-685. doi:10.1080/13607860701529965

83. O'rourke N. Reliability generalization of responses by care providers to the Center for Epidemiologic Studies-Depression Scale. *Educ Psychol Meas*. 2004;64(6):973-990. doi:10.1177/0013164404268668

84. Hakstian AR, Whalen TE. A k-sample significance test for independent alpha coefficients. *Psychometrika*. 1976;41(2):219-231. doi:10.1007/BF02291840

85. Aguayo R, Vargas C, Emilia I, Lozano LM. A meta-analytic reliability generalization study of the Maslach Burnout Inventory. *Int J Clin Health Psychol*. 2011;11(2):343-361.

86. Bonett DG. Sample size requirements for testing and estimating coefficient alpha. *J Educ Behav Stat*. 2002;27(4):335-340. doi:10.3102/10769986027004335

87. Laird NM, Mosteller F. Some statistical methods for combining experimental results. *Int J Technol Assess Health Care*. 1990;6(1):5-30. doi:10.1017/S0266462300008916

88. Raykov T, Marcoulides GA. Meta-Analysis of Scale Reliability Using Latent Variable Modeling. *Struct Equ Model Multidiscip J*. 2013;20(2):338-353. doi:10.1080/10705511.2013.769396

89. Zangaro GA, Soeken KL. Meta-Analysis of the Reliability and Validity of Part B of the Index of Work Satisfaction Across Studies. *J Nurs Meas*. 2005;13(1):7-22. doi:10.1891/jnum.2005.13.1.7

90. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. *J Stat Softw*. 2010;36(1):1-48. doi:10.18637/jss.v036.i03

91. Schwarzer G, Carpenter J, Rücker G. *Meta-Analysis with R*. Springer; 2015.

92. Krizan Z. Synthesizer 1.0: A varying-coefficient meta-analytic tool. *Behav Res Methods*. 2010;42(3):863-870. doi:10.3758/BRM.42.3.863

93. Borenstein M, Hedges LV, Higgins J, Rothstein H. *Comprehensive Meta-Analysis*. Biostat Inc.; 2014.

94. Rosenberg MS, Adams DC, Gurevitch J. *Metawin: Statistical Software for Meta-Analysis with Resampling Tests.* Sinauer Associates; 1997.

95. Cheung MW-L. metaSEM: an R package for meta-analysis using structural equation modeling. *Front Psychol*. 2015;5:1521. doi:10.3389/fpsyg.2014.01521

96. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*. 2016;6(7):e010247. doi:10.1136/bmjopen-2015-010247

97. Charter RA. A Breakdown of Reliability Coefficients by Test Type and Reliability Method, and the Clinical Implications of Low Reliability. *J Gen Psychol*. 2003;130(3):290-304. doi:10.1080/00221300309601160

98. Nunnally JC. *Psychometric Theory*. 3rd ed. McGraw-Hill; 1994.

99. Cicchetti D. Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instrument in Psychology. *Psychol Assess*. 1994;6:284-290. doi:10.1037/1040-3590.6.4.284

100. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *biometrics*. 1977;33:159-174. doi:10.2307/2529310

101. Shankar V, Bangdiwala SI. Observer agreement paradoxes in 2x2 tables: comparison of agreement measures. *BMC Med Res Methodol*. 2014;14(1):100. doi:10.1186/1471-2288-14-100

Table 1. REGEMA Checklist

TITLE		**Yes**	**No**	**Unclear**	**NA**
1. Title	In the title include: (a) the term “reliability generalization” or “meta-analysis” together with some explicit indication to reliability (internal consistency, test-retest, inter- or intra-rater) and (b) the name of the scale or, if more than one scale, the attribute/outcome measure that the scales are assessing.				
ABSTRACT		**Yes**	**No**	**Unclear**	**NA**
2. Abstract	In the abstract explicitly state: (a) that the objective was to carry out a reliability generalization (RG) meta-analysis of one or several scales; (b) eligibility criteria of the studies; (c) data sources with the temporal range covered; (d) types of reliability coefficients analyzed; (e) statistical model applied; (f) main results (e.g., pooled reliability coefficient and 95% CI, moderator variables related to reliability); and (g) main conclusions. In case of space limitation, (b) and (c) criteria can be omitted.				
INTRODUCTION		**Yes**	**No**	**Unclear**	**NA**
3. Background	In the background include: (a) a conceptual definition of the attribute/outcome measure assessed by the scale/s; (b) description of the target population/s to which the scale/s is/are applied and its/their purposes (e.g., screening, clinical diagnosis); (c) a complete description of the scale/s (length, number of categories), including the versions and adaptations to other languages/cultures; and (d) a brief presentation of reliability estimates obtained in previous psychometric studies of the scale/s. Optionally, a brief review of validation studies of the scale/s (e.g., exploratory/confirmatory factor analyses, concurrent/convergent/discriminant validity, responsiveness) could be included.				
4. Objectives	State whether the purpose of the meta-analysis was to obtain a more precise overall reliability coefficient estimate and/or investigate how reliability coefficients vary among different applications of the scales. Optionally, specify whether one objective of the meta-analysis is to estimate the reliability induction rates of the scale/s.				
METHOD		**Yes**	**No**	**Unclear**	**NA**
5. Selection criteria	Specify inclusion criteria: (a) name/s of the scale/s analysed in the RG meta-analysis, as well as the versions and/or adaptations included; (b) geographical and/or cultural restrictions; (c) years considered; (d) language of the paper; (e) publication status; (f) to report any reliability estimate based on the study-specific sample/s; (g) type/s of reliability considered (e.g., internal consistency, temporal stability, inter-				

	/intra rater reliability…); (h) target population/s (e.g., community, clinical, subclinical/analog, university…); and (i) minimum sample size required.				
6. Search strategies	Specify how the studies were located: (a) electronic databases consulted; (b) other formal search procedures (e.g., manual search in specific journals, backward search from references listed in selected studies); and (c) informal search procedures (e.g., internet searches, contacting study authors to identify additional studies). For electronic searches, describe the search strategy, including the keywords used and how they were combined, and the search limits (e.g., fields where the keywords were searched - title, abstract, full-text -, temporal range, language).				
7. Data extraction	Describe the characteristics extracted from the studies, including: (a) sample size/s, mean/s and standard deviation/s of total test scores and subscales (if applicable); (b) sample characteristics (e.g., target population, country, mean age, standard deviation of the age, gender distribution, ethnic distribution, disorder history –mean and SD in years); (c) test version (e.g., adaptation/version, number of items, reporting format –self-report, clinician); (d) methods (e.g., study design, purpose of the study –psychometric versus applied–, quality checklist); (e) extrinsic characteristics (e.g., publication status, researchers' affiliations, funding source).				
8. Reported reliability	Identify the types of reliability coefficients included in the RG meta-analysis: internal consistency (e.g., Cronbach's alpha, KR-21, parallel forms, omega), temporal stability (test-retest), inter- and intra-rater reliability (e.g., intraclass correlation, kappa coefficient). Clearly state that separate meta-analyses were conducted for each type of reliability coefficient. In case of applying a multivariate/MASEM approach, specify the type of statistical information extracted from the studies (i.e., item-item correlation/covariance matrices, factor loadings, etc.).				
9. Estimating the reliability induction and other sources of bias	In case that the meta-analysis intends to estimate the reliability induction, identify the types of reliability induction: induction by omission (no mention of test reliability whatsoever) or reporting induction (vague or precise reporting). Describe how other sources of bias were assessed (e.g. assumptions of the reliability coefficient, adequacy of the measurement model, etc.).				
10. Data extraction of inducing studies	Declare whether characteristics of inducing studies were also extracted or if, on the contrary, only characteristics of studies that reported reliability were extracted.				

11. Reliability of data extraction	Describe how the reliability of data extraction process was appraised: how many coders which agreement coefficients were applied (e.g., kappa coefficient, intraclass correlation), which values were obtained, and how disagreements were dealt with.				
12. Transformation method	State whether or not the reliability coefficients were transformed for the meta-analytic integration. If relevant, specify the transformation methods: Fisher´s Z for correlation coefficients (e.g., test-retest coefficients), Bonett's and Hakstian and Whallen's transformation for internal consistency coefficients (e.g., Cronbach's alpha), reliability index, measurement error (e.g., standard error of measurement), or other (specify).				
13. Statistical model	Describe the statistical model(s) assumed in the meta-analytic integration for estimating the average reliability coefficient and for analysing the influence of moderator variables (e.g. fixed-effect(s), random-effects, mixed-effects, varying-coefficient models, generalized linear models), as well as the analysis framework (frequentist or Bayesian). In case of applying a multivariate/MASEM approach, describe how the item correlation/covariance matrices or factor loadings were synthesized.				
14. Weighting method	Specify the weighting method applied in the meta-analytic integration: unweighted, weighting by sample size, weighting by inverse variance, or other weighting methods.				
15. Heterogeneity assessment	Describe how heterogeneity among reliability coefficients was assessed (e.g., standard deviation, *Q* statistic, I^2 index, between-studies variance, 75% rule of Hunter-Schmidt). If relevant, specify the between-studies variance estimator (DerSimonian and Laird, Maximum Likelihood, Restricted Maximum Likelihood, Empirical Bayes, Paule and Mandel), as well as how confidence intervals, credibility intervals, or prediction intervals were calculated.				
16. Moderator analyses	If relevant, describe how the influence of moderator variables was assessed (e.g., subgroup analyses, meta-regression analyses, correlational analyses).				
17. Additional analyses	Describe other additional analyses accomplished, such as sensitivity analyses (e.g., statistical analyses with transformed and untransformed reliability coefficients, one-to-one deleting of reliability coefficients, assessment of publication bias, reporting biases, and other sources of bias).				
18. Software	Mention the software and version used to carry out the statistical analyses (e.g., metafor in R, Proc MIXED in SAS, Comprehensive Meta-analysis).				
RESULTS		**Yes**	**No**	**Unclear**	**NA**

19. Results of the study selection process	Describe, ideally with a flow chart, the selection process of the studies, specifying the number of studies identified from each search source, excluded studies and reasons why, and the number of studies that reported and induced reliability of test scores. Regarding reliability induction, report induction rates, distinguishing between induction “by omission” and “by report” (see e.g., REGEMA flowchart). Furthermore, it is advisable to compare the reliability induction rates as a function of variables such as publication year, country/continent and study purpose (psychometric vs. applied).				
20. Mean reliability and heterogeneity	Present pooled reliability coefficients and confidence/credibility intervals for the scale (and subscales, if applicable) and for each type of reliability (e.g., internal consistency, temporal stability, inter- and intra-rater agreement). In case of applying any transformation of the reliability coefficients, results should be back-transformed to the original metric to facilitate interpretation. Illustrate the distribution of reliability coefficients with graphical techniques (e.g., forest plots, box plots, stem and leaf displays, histograms) and describe the degree of heterogeneity by one or more heterogeneity measures (see Item 15).				
21. Moderator analyses	For categorical moderators, provide the pooled reliability coefficient, confidence interval and other heterogeneity measures for each category of the moderator. For continuous moderators, include the regression coefficients, standard errors and confidence limits. For both types of moderators, report results of the statistical significance tests, misspecification tests, and proportion of variance accounted for. As a further step, it is advisable to fit a predictive/explanatory model including the most relevant moderator variables.				
22. Sensitivity analyses	Report or describe the results of any sensitivity analyses conducted (see Item 17).				
23. Comparison of inducing and reporting studies	If performed, present the results of comparing the characteristics of inducing and reporting studies (e.g., sociodemographic and clinical characteristics of the samples).				
24. Data set	Tabulate the characteristics of the individual studies that reported reliability (see Item 7). Tables can be presented as appendices or supplementary files. In addition, list of all studies included in the RG meta-analysis, either in the reference section or as a supplementary file.				
DISCUSSION		**Yes**	**No**	**Unclear**	**NA**
25. Summary of results	Present the main results, such as mean reliability exhibited by the scale/test and moderators of the reliability coefficients. If available, discuss the results in the light of previous evidence.				
26. Limitations	Discuss the limitations of the meta-analysis. Include an explicit statement of the reliability induction rates and the extent to which inducing and reporting studies are comparable in terms of samples characteristics.				
27. Implications for practice	Provide guidelines for professional practice regarding the usefulness of the scale/test in different settings and target populations.				

28. Implications for future research	Include recommendations for researchers regarding the conditions under which the scale/test should be applied.				
FUNDING		**Yes**	**No**	**Unclear**	**NA**
29. Funding	State the financial sources of the meta-analysis, as well as potential conflict of interests of the authors.				
PROTOCOL		**Yes**	**No**	**Unclear**	**NA**
30. Protocol	State whether a protocol of the meta-analysis was previously published or made accessible in some web-site (e.g., in Prospero).				

Note. NA: Not Applicable.

Table 2. Results of the analysis of the inter-coder reliability of the REGEMA checklist.

Item	% agreement	Kappa
Item 1: Title	97	.875
Item 2: Abstract	87	.627
Introduction:		
Item 3: Background	87	.742
Item 4:Objectives	95	.643
Method:		
Item 5: Selection criteria	80	.595
Item 6: Search strategies	92	.848
Item 7: Data extraction	87	.677
Item 8: Reported reliability	90	.279
Item 9: Reliability induction and other sources of bias	92	.842
Item 10: Data extraction of inducing studies	97	.655
Item 11: Reliability of data extraction	95	.900
Item 12: Transformation method	97	.950
Item 13: Statistical model	95	.896
Item 14: Weighting method	90	.842
Item 15: Heterogeneity assessment	87	.741
Item 16: Moderator analyses	93	.899
Item 17: Additional analyses	90	.749
Item 18: Software	97	.947
Results:		
Item 19: Results of the study selection process	100	1
Item 20: Mean reliability and heterogeneity	90	.799
Item 21: Moderator analyses	100	1
Item 22: Sensitivity analyses	90	.792
Item 23: Comparison of inducing and reporting studies	100	1
Item 24: Data set	90	.737
Discussion:		
Item 25: Summary of results	97	.655
Item 26: Limitations	95	.875
Item 27: Implications for practice	85	.701
Item 28: Implications for future research	87	.474
Item 29: Funding	92	.826
Item 30: Protocol[a]	100	--

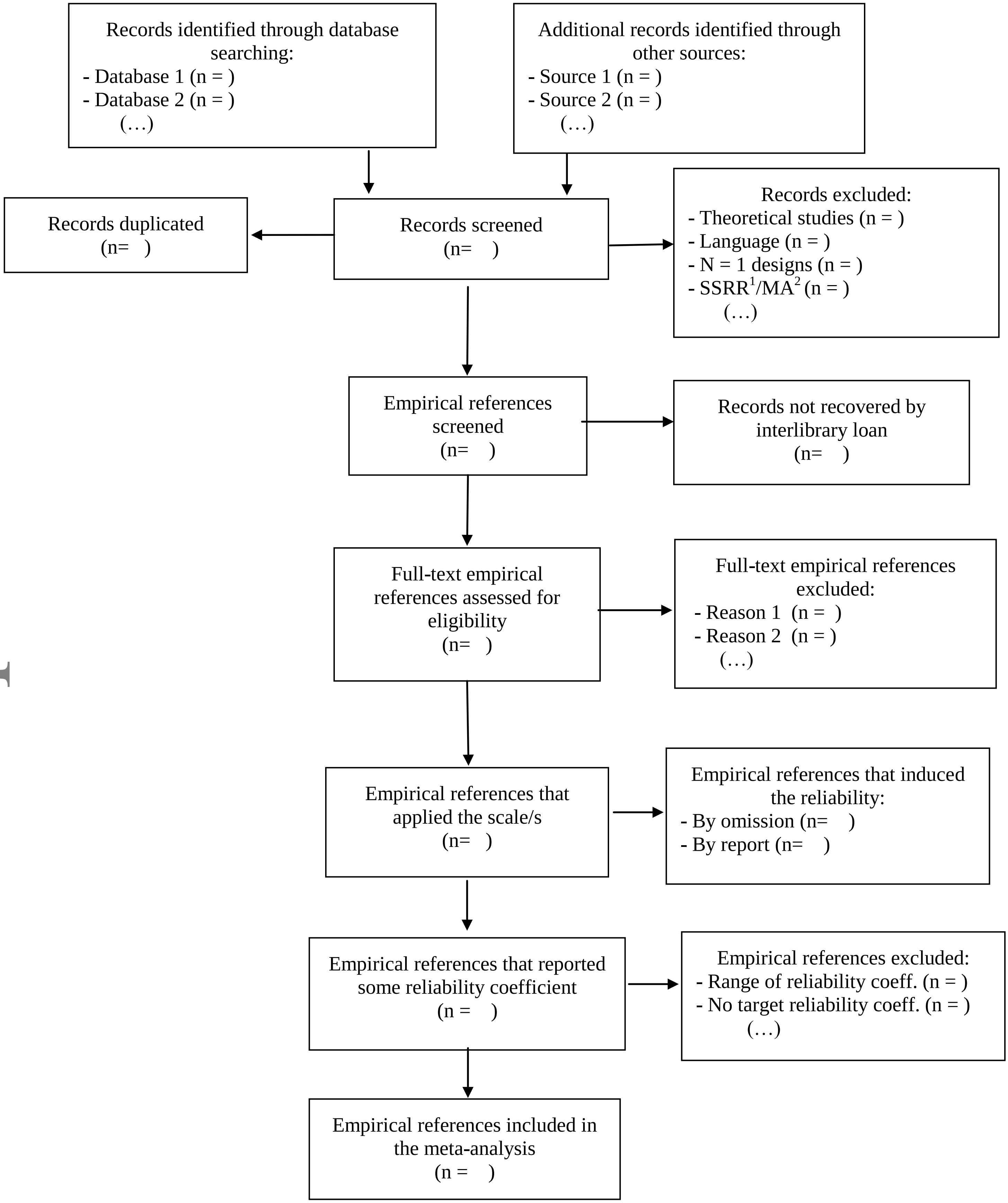
[a] It was not possible to calculate Cohen’s kappa because the dependent variable was a constant.

Table 3. Results of the analysis of the compliance with the REGEMA checklist of 40 RG meta-analyses in Psychology.

Item	% of compliance
Item 1: Title	90
Item 2: Abstract	90
Introduction:	
Item 3: Background	**45**
Item 4:Objectives	95
Method:	
Item 5: Selection criteria	57
Item 6: Search strategies	55
Item 7: Data extraction	**27**
Item 8: Reported reliability	92
Item 9: Reliability induction and other sources of bias	**37**
Item 10: Data extraction of inducing studies	**2**
Item 11: Reliability of data extraction	**47**
Item 12: Transformation method	50
Item 13: Statistical model	60
Item 14: Weighting method	62
Item 15: Heterogeneity assessment	**40**
Item 16: Moderator analyses	52
Item 17: Additional analyses	**30**
Item 18: Software	**37**
Results:	
Item 19: Results of the study selection process	**22**
Item 20: Mean reliability and heterogeneity	55
Item 21: Moderator analyses	90
Item 22: Sensitivity analyses	**42**
Item 23: Comparison of inducing and reporting studies	**2**
Item 24: Data set	30
Discussion:	
Item 25: Summary of results	95
Item 26: Limitations	72
Item 27: Implications for practice	50
Item 28: Implications for future research	87
Item 29: Funding	**35**
Item 30: Protocol	**0**

Items in boldface presented percentages of compliance under 50%.

REGEMA flow diagram


Records identified through database searching:
- Database 1 (n =)
- Database 2 (n =)
(…)
Additional records identified through other sources:
- Source 1 (n =)
- Source 2 (n =)
(…)
Records duplicated (n=)
Records screened (n=)
Records excluded:
- Theoretical studies (n =)
- Language (n =)
- N = 1 designs (n =)
- SSRR1/MA2 (n =)
(…)
Empirical references screened (n=)
Records not recovered by interlibrary loan (n=)
Full-text empirical references assessed for eligibility (n=)
Full-text empirical references excluded:
- Reason 1 (n =)
- Reason 2 (n =)
(…)
Empirical references that applied the scale/s (n=)
Empirical references that induced the reliability:
- By omission (n=)
- By report (n=)
Empirical references that reported some reliability coefficient (n =)
Empirical references excluded:
- Range of reliability coeff. (n =)
- No target reliability coeff. (n =)
(…)
Empirical references included in the meta-analysis (n =)

Note. [1] SSRR = Systematic Reviews. [2] MA = Meta-analyses.