UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

Statistical Methodology of Reliability Generalization Meta-Analysis.

Metodología estadística del Meta-Análisis de Generalización de la Fiabilidad.

D. Carmen López Ibáñez

2023

UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

Statistical Methodology of Reliability Generalization Meta-Analysis.

Metodología estadística del Meta-Análisis de Generalización de la Fiabilidad

Autor: D. Carmen López-Ibáñez

Director/es: D. Julio Sánchez-Meca

UNIVERSIDAD DE
MURCIA

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD
DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

Aprobado por la Comisión General de Doctorado el 19-10-2022

D./Dña. Carmen López Ibáñez

doctorando del Programa de Doctorado en

Psicología

de la Escuela Internacional de Doctorado de la Universidad Murcia, como autor/a de la tesis presentada para la obtención del título de Doctor y titulada:

Statistical Methodology of Reliability Generalization Meta-Analysis.
Metodología estadística del Meta-Análisis de Generalización de la Fiabilidad

y dirigida por,

D./Dña. Julio Sánchez Meca

D./Dña.

D./Dña.

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita, cuando se han utilizado sus resultados o publicaciones.

Si la tesis hubiera sido autorizada como tesis por compendio de publicaciones o incluyese 1 o 2 publicaciones (como prevé el artículo 29.8 del reglamento), declarar que cuenta con:

- *La aceptación por escrito de los coautores de las publicaciones de que el doctorando las presente como parte de la tesis.*
- *En su caso, la renuncia por escrito de los coautores no doctores de dichos trabajos a presentarlos como parte de otras tesis doctorales en la Universidad de Murcia o en cualquier otra universidad.*

Del mismo modo, asumo ante la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada, en caso de plagio, de conformidad con el ordenamiento jurídico vigente.

En Murcia, a 28 de septiembre de 2023

LOPEZ IBAÑEZ
CARMEN -
48744672J

Firmado digitalmente por
LOPEZ IBAÑEZ CARMEN -
48744672J
Fecha: 2023.09.28 10:37:03
+02'00'

Fdo.: Carmen López Ibáñez

Esta DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD debe ser insertada en la primera página de la tesis presentada para la obtención del título de Doctor.

Información básica sobre protección de sus datos personales aportados	
Responsable:	Universidad de Murcia. Avenida teniente Flomesta, 5. Edificio de la Convalecencia. 30003; Murcia. Delegado de Protección de Datos: dpd@um.es
Legitimación:	La Universidad de Murcia se encuentra legitimada para el tratamiento de sus datos por ser necesario para el cumplimiento de una obligación legal aplicable al responsable del tratamiento. art. 6.1.c) del Reglamento General de Protección de Datos
Finalidad:	Gestionar su declaración de autoría y originalidad
Destinatarios:	No se prevén comunicaciones de datos
Derechos:	Los interesados pueden ejercer sus derechos de acceso, rectificación, cancelación, oposición, limitación del tratamiento, olvido y portabilidad a través del procedimiento establecido a tal efecto en el Registro Electrónico o mediante la presentación de la correspondiente solicitud en las Oficinas de Asistencia en Materia de Registro de la Universidad de Murcia

INDEX

Resumen

Una parte fundamental de las ciencias sociales y de la salud, como la Psicología, es cuantificar las capacidades, los rasgos o los atributos psicológicos con los que cuentan las personas. Para que esta cuantificación sea verdaderamente representativa y pueda ser útil, es necesario que el proceso de medición sea preciso y se realice a través de instrumentos de medida bien construidos y contrastados, es decir, estos instrumentos deben cumplir una serie de mínimos que arrojen resultados confiables. El instrumento en cuestión debe contar con un protocolo de administración donde se determine la forma correcta de aplicarlo, incluyendo el rango de aplicabilidad y estableciendo el objetivo de dicho instrumento y el tipo de población al que va dirigido. También es importante que la medición sea *fiable*, es decir, que las diferentes aplicaciones de este arrojen resultados concordantes. Otro aspecto fundamental sobre el que debe construirse todo el proceso de medición es que las inferencias establecidas sobre el atributo o el rasgo que se pretende medir sean *válidas*, implicando una fundamentación teórica subyacente al instrumento bien consolidada (Abad et al., 2011).

Bajo estas premisas y con la intención de dotar a la medición psicológica de rigor científico, a lo largo del siglo XX se desarrolló y profundizó la conocida Teoría Clásica de los Tests (TCT), siendo Charles Spearman durante los primeros 15 años su principal precursor (Spearman, 1904, 1907, 1913). El objetivo central de esta teoría es establecer un modelo estadístico que permita realizar asociaciones entre el nivel de rasgo latente o verdadero de la persona, y el nivel de rasgo empírico o medido por el instrumento. Este modelo asume que las puntuaciones de ambos niveles no van a ser idénticas, ya que presupone que existen multitud de factores imposibles de cuantificar que van a afectar a

la medida en el momento en el que se está evaluando a la persona. Es por ello que el modelo incluye en su formulación un componente de error asociado a todo este proceso de medición (Muñiz, 2018).

Puesto que la puntuación verdadera de una persona es completamente desconocida e imposible de cuantificar, no podemos establecer relaciones directas entre la puntuación observada y la verdadera. Lo que sí podemos conocer es la relación entre dos formas paralelas de un mismo test diseñadas para evaluar un determinado rasgo. Estas formas paralelas por definición deben cumplir dos condiciones: (1) la puntuación verdadera en ambas formas debe ser la misma, y (2) la varianza de los errores de medida también debe ser la misma en las dos formas. Es decir, ambas formas deben medir exactamente lo mismo y con la misma precisión (Abad et al., 2011). La correlación entre ambas formas se denomina *coeficiente de fiabilidad* y nos permite inferir el grado de precisión de cada una de ellas y establecer la parte de varianza de las puntuaciones observadas que se deben a la variabilidad de las puntuaciones verdaderas. En este contexto, entendemos *fiabilidad* como una propiedad psicométrica que nos indica la replicabilidad de la medida.

Esta replicabilidad cuantificada a través del coeficiente de fiabilidad puede expresarse de tres maneras distintas: en primer lugar, puede representar el grado de equivalencia entre diferentes formas del test, es decir, la medida será la misma cuando se mide el mismo rasgo con pruebas equivalentes. En segundo lugar, puede entenderse como estable en el tiempo, indicando que la medida es la misma en dos momentos temporales diferentes. Y por último, puede significar el grado de consistencia, cuando la replicabilidad radica en medir lo mismo con diferentes partes de un test (Abad et al., 2011). En función de cada interpretación de la replicabilidad de la medida, pueden extraerse y calcularse diferentes tipos de coeficientes de fiabilidad.

En la práctica, uno de los coeficientes de fiabilidad más utilizados es el coeficiente alfa de Cronbach (Cronbach, 1951), perteneciente al grupo de los coeficientes de consistencia interna, evaluando la concordancia de las puntuaciones entre los ítems. Este coeficiente se considera el límite inferior del coeficiente verdadero (Abad et al., 2011). Alfa está fuertemente influenciado por dos factores: por el grado de correlación promedio entre los ítems del test, y por el número de ítems que componen el test. Además, el rango de valores de este coeficiente oscila entre 0 y 1.

Cabe señalar que la fiabilidad es propia de las puntuaciones, en ningún caso esta propiedad psicométrica hace referencia al instrumento de medida (Crocker & Algina, 1986; Gronlund & Linn, 1990; Thompson & Vacha-Haase, 2000; Traub, 1994). Para poder generalizar el valor obtenido al instrumento en sí mismo, resulta imprescindible aplicar herramientas estadísticas que nos permitan calcular un valor promedio utilizando todas las aplicaciones previas de ese instrumento. La mejor herramienta para ese fin, es decir, que sea capaz de recoger y sintetizar toda la evidencia cuantitativa es el meta-análisis. El meta-análisis es el conjunto de técnicas y procedimientos cuantitativos aplicados a la síntesis de la evidencia que recogen las revisiones sistemáticas. Debido al aumento exponencial de los estudios científicos y de los resultados empíricos, aumenta la necesidad de contar con herramientas que integren y proporcionen conclusiones rigurosas sobre los resultados arrojados por las investigaciones científicas (Botella & Sánchez-Meca, 2015; Cooper et al., 2019; Schmid et al., 2020). Es especialmente interesante este procedimiento ya que los resultados individuales de los estudios -primarios- suelen reportar conclusiones diferentes o incluso, contradictorias. La integración de todos los resultados experimentales en un mismo campo de estudio proporcionan una visión más globalizada y realista del efecto verdadero, y se considera un elemento imprescindible para la construcción del conocimiento (Botella & Sánchez-Meca, 2015).

En 1998, Vacha-Haase implementó esta metodología al estudio de la fiabilidad, denominándolo **Meta-Análisis de Generalización de la Fiabilidad (GF)**. De esta forma, los tamaños del efecto de los estudios aquí son los coeficientes de fiabilidad obtenidos en cada una de las aplicaciones del instrumento. Esas múltiples aplicaciones nos permiten, por otro lado, estudiar la influencia de variables que pueden estar contribuyendo a las fluctuaciones del coeficiente entre los estudios primarios. Al contrario de lo que ocurre en otros tipos de meta-análisis, en un meta-análisis de generalización de la fiabilidad se espera que el efecto de esas variables moderadoras sea limitado o insignificante, ya que esa invariabilidad aporta robustez a los constructos sobre los que se cimienta el instrumento y aporta evidencia de la validez de las puntuaciones obtenidas (Botella & Sánchez-Meca, 2015).

Un aspecto fundamental de esta metodología es su flexibilidad y la ausencia de un protocolo estricto para implementarla. Cada investigador debe tomar las decisiones estadísticas oportunas en función de los datos con los que cuente y la capacidad de generalización de los resultados. Encontramos tres decisiones fundamentales: seleccionar la transformación que se aplicará a los coeficientes de fiabilidad, si se aplicara alguna; establecer qué modelo estadístico se asumirá; y determinar cuál será el método de ponderación de este.

Respecto a la primera decisión, las transformaciones de los coeficientes más frecuentemente aplicadas en los meta-análisis GF son la transformación Z de Fisher, la transformación de Hakstian y Whalen (1976) y la propuesta por Bonett (2002). Se ha observado que tanto Hakstian-Whalen como Bonett normalizan la distribución y, en el caso de Bonett, estabiliza las varianzas. A pesar de que la Z de Fisher es una de las transformaciones más utilizadas, no sería correcto utilizarla para coeficientes de consistencia interna, sino para aquellos coeficientes que se basen en la correlación de Pearson.

En segundo lugar, la segunda decisión que se debe tomar radica en la elección del modelo estadístico, principalmente el modelo de Efecto-Fijo, el modelo de Efectos Aleatorios, el modelo de Coeficientes Variables y el método OLS *(Ordinary Least Squares)*. Aunque es un modelo incorrecto, este último modelo es el más utilizado en el campo. Este modelo consiste en la aplicación de métodos estadísticos convencionales como el cálculo de una media no ponderada de los coeficientes, estimar su varianza muestral y construir un intervalo de confianza del 95% como si las estimaciones de fiabilidad pertenecieran a una única muestra de participantes. Obviar este modelo del trabajo sería obviar gran parte de los meta-análisis GF que hay publicados en la actualidad.

Por otro lado, el modelo de Efecto Fijo asume que los coeficientes de fiabilidad reportados en los estudios están estimando un parámetro poblacional común, entendiendo así que la única fuente de variabilidad entre diferentes estimaciones de fiabilidad se debe al error muestral. El modelo de Coeficientes Variables asume que cada coeficiente de fiabilidad individual está estimando un parámetro poblacional diferente, aunque, al contrario que el modelo de Efectos Aleatorios, este modelo no asume que los coeficientes paramétricos sean una muestra representativa de una superpoblación de potenciales coeficientes de fiabilidad. Tanto el modelo de Efecto Fijo como el modelo de Coeficientes Variables (CV) no buscan generalizar sus resultados más allá de la población de estudios que presenten idénticas características a los incluidos en el estudio. Concretamente, si se sospecha de la existencia de heterogeneidad, estaría completamente desaconsejado el modelo de Efecto Fijo.

Por otro lado, el modelo de Efectos Aleatorios, en cualquiera de sus variantes, determina que la variabilidad encontrada no se explica únicamente por la varianza muestral. Al igual que en el caso del modelo CV, este modelo asume que cada coeficiente de fiabilidad individual está estimando un parámetro poblacional diferente, sin embargo, en este modelo se entiende que esos parámetros constituyen una muestra representativa de una potencial distribución de coeficientes de fiabilidad paramétricos. Este modelo incorpora dos fuentes de variabilidad: la varianza intraestudio, debida al error de muestreo de participantes dentro de cada muestra, y la varianza interestudios, debida al error de muestreo de los coeficientes de fiabilidad verdaderos a partir de una superpoblación de coeficientes de fiabilidad. Este modelo sería el indicado si pretendemos generalizar los resultados del meta-análisis a cualquier estudio, sean cuales sean sus características.

En este punto es donde se tiene que decidir el tipo de ponderación que se va a aplicar. Existen dos modelos alternativos al clásico modelo de Efectos Aleatorios que pondera por la inversa del peso. Por un lado, encontramos el modelo de Efectos Aleatorios de Schmidt y Hunter (2015) que aplica una ponderación por el tamaño muestral de cada estudio. Por el otro lado, el modelo de Efectos Aleatorios mejorado de Hartung y Knapp (2001) cambia la forma de estimar la varianza muestral y asume una distribución t de Student con *k*-1 grados de libertad, siendo *k* el número de estudios. Este modelo tiene en cuenta la incertidumbre en la estimación de la varianza entre estudios τ^2, por lo que ofrece un mejor ajuste que los modelos de Efectos Aleatorios explicados anteriormente.

Teniendo en cuenta la literatura previa del campo, y las incompatibilidades teóricas entre estrategias, la cifra de combinaciones generadas se sitúa en 13 para el cálculo del coeficiente promedio y 18 para el cálculo de su intervalo de confianza. El primer estudio de esta tesis (Capítulo 2) tiene como objetivo la comparación estadística de los resultados generados en cada una de las combinaciones, determinando si diferentes decisiones pueden dar lugar a diferentes conclusiones.

Aunque la naturaleza de esta metodología permite utilizar cualquier estrategia analítica, resulta imprescindible que esta se reporte precisa y minuciosamente, especificando cada una de las decisiones que se han tomado para obtener dichos resultados. De esa forma garantizamos que cualquier investigador pueda reproducirlos o replicarlos. El segundo estudio de esta tesis (Capítulo 3) presenta un trabajo de reproducción de los meta-análisis de generalización de la fiabilidad publicados en revistas científicas. Este estudio, además, presenta una revisión del reporte de las categorías indispensables para reproducir los resultados de cada meta-análisis.

Los meta-análisis de generalización de la fiabilidad también presentan desventajas en su aplicación. Una de ellas radica en el hecho de que no se tienen en cuenta las posibles relaciones de dependencia que pudieran surgir. Cuando un test tiene una estructura multidimensional con varias subescalas, todas formando parte de un mismo constructo psicológico y, además, ese test se está aplicando a diferentes grupos dentro del mismo estudio científico, esta dependencia puede aparecer. Esto significa que un tamaño del efecto observado nos puede estar dando información sobre la dirección o el grado de desviación de otro efecto más allá de lo esperado por el modelo (Assink & Wibbelink, 2016; Van den Noortgate et al., 2013). Tradicionalmente para acabar con estas redes de dependencia, lo que se ha propuesto es dividir cada subescala o grupo muestral en meta-análisis independientes. Esta estrategia no es la más adecuada, ya que los resultados que se obtienen son menos precisos y los análisis estadísticos pierden potencia (Assink & Wibbelink, 2016; Van den Noortgate et al., 2013).

Una alternativa es tratar de modelar la dependencia aplicando modelos multinivel. Como no es necesario que todos los estudios primarios reporten exactamente la misma información, una de las ventajas de este modelo es que pueden utilizarse todos los tamaños del efecto relevantes que aporten los estudios. Además, posee una estructura muy flexible, por lo que puede adaptarse a cualquier tipo de datos. Otra ventaja de estos modelos es que los datos automáticamente se ajustan a la estructura jerárquica del análisis (Van den Noortgate et al., 2013). El modelo multinivel que más se utiliza y que parece que mejor funciona es el modelo estructurado en tres niveles (Assink & Wibbelink, 2016; Van den Noortgate et al., 2013, 2015). Esta estructura distribuye los componentes de la varianza en tres niveles: el primero, la varianza muestral de los tamaños del efecto (en este caso, de los coeficientes de fiabilidad); el segundo nivel hace referencia a la varianza entre efectos dentro del mismo estudio, y, por último, el tercer nivel incorpora la varianza entre los estudios.

Teniendo en cuenta el tipo de estudios que componen un meta-análisis GF no parece muy descabellado que aparezcan relaciones de dependencia. Sobre todo, si tenemos en cuenta que habitualmente un instrumento de medida se administra a varios grupos dentro de un mismo estudio, o que los instrumentos están compuestos por varias subescalas que miden diferentes componentes de un mismo rasgo psicológico. Es por ello que el tercer estudio de esta disertación (Capítulo 4) tiene por objetivo comprobar si los resultados difieren estadísticamente cuando se realiza un meta-análisis GF desde el punto de vista convencional, que rompe las posibles redes de dependencia realizando meta-análisis independientes, y los resultados de un meta-análisis multinivel de tres niveles, que modela dicha dependencia dentro de un único análisis. Además, se han tenido en cuenta otros factores como la transformación de los coeficientes o el método de cálculo del intervalo de confianza del coeficiente de fiabilidad promedio.

En resumen, esta tesis tiene tres objetivos fundamentales: determinar si las decisiones estadísticas que se toman a lo largo del proceso de realización de un meta-análisis de generalización de la fiabilidad condicionan los resultados obtenidos; comprobar si los resultados distan entre un meta-análisis que abole las posibles relaciones de dependencia y uno que las integra y modela; y, por último, comprobar el grado de reproducibilidad de este tipo de estudios, así como el grado de transparencia y reporte de la información fundamental para repetir los análisis.

Chapter 1

Introduction

1.1 Classical Test Theory

Quantifying the psychological capacities, traits or attributes of individuals is a fundamental part of social and health sciences such as psychology. Personality traits, intellectual level, or the severity of symptoms of a certain disorder are some examples. To ensure that this quantification is truly representative and useful, the assessment process must be precise and be carried out using well-constructed and contrasted measurement instruments, that is, these instruments must meet a series of minimum requirements that yield reliable results. The instrument concerned must have an administration protocol that determines the correct way to apply it. This protocol must also have a range of applicability, specifying the purpose of the instrument and its target population. Another important point is that the measurement should be *reliable,* that is to say, that the different applications of the instrument should yield consistent results. A further essential element on which the whole assessment process must be built is that the inferences drawn about the attribute or trait to be measured are *valid*, implying a well-established theoretical foundation underlying the instrument (Abad et al., 2011).

Under these premises and in order to provide psychological measurement scientific rigour, throughout the 20th century the well-known Classical Test Theory (CTT) was developed and deepened, with Charles Spearman being its main precursor during the first 15 years (Spearman, 1904, 1907, 1913). The main focus of this theory is to establish a statistical model that allows to make associations between the latent or true trait level of the person, and the empirical trait level measured by the instrument. This model assumes that the scores at both levels will not be identical, due to the fact that there are a multitude of unquantifiable factors that affect the measure at the moment the person is being assessed. For this reason, the model includes in its formulation an error component associated with the whole measurement process (Muñiz, 2018). Mathematically, the model is expressed as:

$$X = V + e \qquad [1.1]$$

where *X* refers to the observed score, *V* to the true score, and *e* to the error associated with the measurement process.

Since a person's true score is completely unknown and impossible to quantify, no direct relationship can be established between the observed and true score. However, the relationship between two parallel forms of the same test designed to assess a certain trait can be known. These parallel forms, by definition, must meet two conditions: (1) the true score in both forms must be the same, and (2) the variance of the measurement errors must also be the same in both forms. That is, both forms must measure exactly the same and with the same precision (Abad et al., 2011). The correlation between the two forms is called the *reliability coefficient* and allows to infer the degree of precision of each of them and to establish the part of the variance of the observed scores that is due to the variability of the true scores. In this context, *reliability* is understood as a psychometric property that indicates the replicability of the measure.

This replicability quantified through the reliability coefficient can be understood in three main ways: first, it can be understood as the degree of equivalence between different forms of the test. That is, replicability indicates that the measure should be the same when the same trait is measured with equivalent tests. Secondly, it can be understood as stable over time, if replicability indicates that the measure is the same at two different points in time. And finally, it can be understood as the degree of consistency, when the replicability lies in measuring the same thing with different parts within the same test (Abad et al., 2011). Depending on each interpretation of the replicability of the measure, different types of reliability coefficients can be extracted and calculated.

In practice, one of the most widely employed is Cronbach's alpha coefficient (Cronbach, 1951), belonging to the group of internal consistency coefficients, which assesses the concordance of scores between items. Mathematically, the alpha coefficient is expressed as:

$$\alpha = \frac{J}{J-1} \frac{\sum_{j \neq j'} \sigma_{X_j X_{j'}}}{\sigma_X^2} \qquad [1.2]$$

where J is the number of items, $\sum_{j \neq j'} \sigma_{X_j X_{j'}}$ is the sum of the empirical covariances between items, and σ_X^2 is the variance of the empirical test scores. This coefficient is considered to be the lower limit of the true coefficient, that is, it will always be higher than the coefficient obtained with the formula $[\alpha \leq \sigma_V^2 / \sigma_X^2]$ (Abad et al., 2011). Alpha is strongly influenced by two factors: the average degree of correlation between test items, and the number of items in the test. The range of this coefficient is between 0 and 1.

1.2 Reliability Generalization Meta-Analysis

Note that reliability is a psychometric property of the scores and not of the instrument (Crocker & Algina, 1986; Gronlund & Linn, 1990; Thompson & Vacha-Haase, 2000; Traub, 1994). In 1998, Vacha-Haase developed the concept of "reliability generalization" in order to synthesize all the results of the reliability coefficient calculation of multiple applications of the same test into an average coefficient for the instrument itself. For this purpose, she developed this concept under the umbrella of the best and most complete tool for synthesis of evidence: meta-analysis.

Meta-analysis is a set of quantitative techniques and procedures applied to the synthesis of evidence collected in systematic reviews. Given the rise of science and empirical results, there is a growing need for tools that integrate and provide rigorous conclusions on the results of scientific research (Botella & Sánchez-Meca, 2015; Cooper et al., 2019; Schmid et al., 2020). These individual results of the -primary- studies often report different or even contradictory conclusions. Therefore, the integration of all of them, as well as the analysis of the different variables that may affect these results, can give us a more comprehensive view about what is happening under this topic. In fact, the integration of research that shares the same focus of study is considered an essential element in the construction of knowledge (Botella & Sánchez-Meca, 2015). Currently it is the most frequently employed and most reputable technique in the field of evidence synthesis. Glass (1976) is considered the first published meta-analysis, and it was carried out in the field of effectiveness evaluation in psychotherapy.

Applying this technique to psychometric analyses of instrument reliability, the reliability generalization meta-analysis collects all the reliability coefficient estimates obtained in each application of the instrument and provides a combined estimate for all of them. These multiple applications allow, in addition, to study the influence of variables that may be contributing to fluctuations in the coefficient between primary studies. However, contrary to what happens in other types of meta-analyses, in a reliability generalization meta-analysis is expected that the effect of these moderating variables is limited or insignificant, since this invariance provides robustness to the constructs underlying the instrument and provides evidence of the validity of the scores obtained (Botella & Sánchez-Meca, 2015).

A peculiarity of this type of meta-analysis resides in the fact that when Vacha-Haase (1998) proposed this methodology, she did not include a single analytical strategy to perform it, but left it to the discretion of each researcher to choose the analyses that best fit the data of the study or the generalization of the conclusions that are intended to be drawn. It should be noted that, although it is possible to use any analytical strategy within this type of meta-analysis, it is essential that this be reported precisely and thoroughly in the papers. In this way it can be ensured that any researcher can reproduce or replicate the results.

In recent years, the reproducibility and replicability of the psychological research has become an important topic (McNutt, 2014; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). Meta-analyses are not free from these problems; therefore, efforts to investigate factors that may affect the reproducibility of meta-analyses are warranted (Lakens et al., 2016).

1.2.1 Phases of Reliability Generalization Meta-Analysis

The procedure for conducting a reliability generalization meta-analysis follows the same steps as a conventional effect size meta-analysis: (a) definition of the research question, (b) literature search, (c) coding of studies and data extraction, (d) statistical analysis and interpretation of results, and (e) publication.

(A) Definition of the research question

The first step for conducting a reliability generalization meta-analysis, as in any empirical study, is to define the objectives of the research itself. To do this, it is necessary to determine what is to be analysed and to define the research problem theoretically and operationally. Expressing it with an example, we could establish as an objective the study of a specific measurement tool that assesses symptoms of obsessive-compulsive disorder, setting out the meta-analysis on the reliability of the instrument..

(B) Literature search

Once the main objective of the research is known, the search strategy must be established. This strategy has to be sufficiently precise and thorough to collect as many published and unpublished studies and papers as possible that have applied the instrument to be meta-analyzed.

At this point, the search engines and databases for the electronic search of the primary studies are determined. Also at this stage, the characteristics of this search will be defined, such as the keywords, the time range, or the inclusion criteria of the studies. Regarding the latter, what is defined are all those characteristics that the studies must present in order to form part of the study (the type of reliability coefficient used, whether it has been calculated with the study's own sample, whether the study is a clinical or population-based sample, etc.). Exclusion criteria are also established at this point.

(C) Coding of studies and data extraction

Following the filtering of the studies that appear in the search according to the inclusion and exclusion criteria, the coding criterion is established for the articles that include all the variables that the author considers to be of interest for the work. It is also recommended that a coding manual be drawn up specifying each of the variables and their coding code.

From this, a database is set up to record all the data extracted from the primary studies according to the variables highlighted.

(D) Statistical analysis and interpretation of results

Statistical analysis involves selecting the analytical strategy to be carried out. This strategy must be chosen according to the type of data we are analyzing and the degree of generalization that is intended to be achieved with the conclusions of the study. The statistical model on which the meta-analysis is to be performed, whether the coefficients are to be transformed or not, and the weighting to be applied must be chosen. In the vast majority of studies, the average reliability coefficient and its confidence interval are calculated at this point, and different heterogeneity estimators are also used (I^2 and Cochrane’s Q indices, prediction intervals, etc.). Also at this stage, the quantitative and qualitative moderator variables that can be analyzed are established to determine the degree of influence they have on the reliability coefficient (gender of the sample, type of population, mean age, mean test scores, etc.).

(E) Publication

Empirical research should share its results with the scientific community and promote the advancement of knowledge to be accessible to any researcher. One of the ways to carry out this diffusion of new knowledge is the publication of research as scientific articles in impact journals. In addition, in order to allow other researchers to replicate or reproduce the results, as well as to assess the methodological quality of the research, it is important that the article is written with full transparency and clarity.

For this purpose, guidelines have been developed to help researchers to correctly report all relevant information in their works. For reliability generalization meta-analysis, the guideline developed specifically for the correct reporting of this type of study is The REGEMA checklist (Sánchez-Meca et al., 2021).

1.2.2 Analytical strategies for conducting an RG meta-analysis

Sánchez-Meca et al. (2012) identified a variety of methods to statistically integrate reliability coefficients. Differences among the methods refer to the statistical model assumed, whether reliability estimates must be transformed to normalize their distribution and stabilize their variances, and whether to weight the reliability estimates when they are statistically integrated.

The result of these different methods has led to large variability in statistical methods applied in RG meta-analyses. An issue not yet investigated is whether the choice of different statistical methods can lead to substantial changes in RG meta-analysis results. If different methods applied to the same RG meta-analysis have an impact in their results, then their conclusions will be conditioned by the methods applied. In addition, the results of RG meta-analyses applying different methods cannot be compared. As a consequence, the diversity of methods to statistically integrate reliability coefficients in RG meta-analyses has consequences in the comparability of their results, as well as in their reproducibility by other researchers.

1.2.2.1 Transformation of coefficients

When conducting an RG meta-analysis of reliability coefficients, the meta-analyst must decide whether coefficients should be transformed to normalize distribution and stabilize variances. Some authors advise against transforming reliability coefficients

(Henson & Thompson, 2002; Mason et al., 2007), whereas others are in favor (Rodriguez & Maeda, 2006; Sánchez-Meca et al., 2012). Not all transformations are recommended for all types of reliability coefficients. For example, for indices based on Pearson's correlation coefficient, which ranges between -1 and +1, such as split-half reliability coefficients and test-retest reliability, the most appropriate transformation would be Fisher's Z transformation. However, for those coefficients ranging between 0 and 1 (Cronbach's alpha, McDonald's omega or split-rater reliability, among others), it would be theoretically more correct to use the transformations proposed by Bonett (2002) or Hakstian and Whalen (1976) Hakstian and Whalen (1976).

The difference between the application of the different coefficient transformations as well as their implications will be discussed further in Chapter 2.

1.2.2.2 Statistical models

Another important decision in a meta-analysis is choosing the statistical model under which the statistical analyses will be accomplished. Fixed-effect (FE) and random effects (RE) models are the two most commonly used statistical models in meta-analysis. Under an FE model the meta-analyst assumes that the reliability coefficients reported in the studies are estimating a common population parameter, so that the only variability source among reliability estimates is due to sampling error.

When an RE model is assumed, the meta-analyst is then acknowledging that the reliability estimates exhibit more variability than sampling error can explain. The extra heterogeneity is due to the fact that each reliability coefficient is estimating a different parameter, these parameters constituting a representative sample of a distribution of potential parametric reliability coefficients. RE models take into account two variability sources: within-study variance (i.e., the same as in the FE model) due to sampling of participants in each sample and between-studies variance owing to sampling of true reliability coefficients from a super-population of reliability coefficients.

Another model that seems to be situated between the two is known as the varying-coefficient (VC) model was proposed in the meta-analytic arena by Laird and Mosteller (1990) and advocated by Bonett (2010) to be applied in RG meta-analysis. Like the RE model, VC assumes that each individual reliability coefficient is estimating a different population parameter, but contrary to the RE model, VC does not assume that the parametric reliability coefficients are a representative sample of a larger population of potential reliability coefficients.

Finally, a model that is certainly controversial but widely used in this field is the ordinary least squares (OLS) method. OLS method consists of applying conventional statistical methods, that is, to calculate an unweighted mean of reliability coefficients, to estimate its sampling variance, and to construct a 95% confidence interval as if the reliability estimates were single data from a sample of participants. Although the OLS method can be thought of as an FE model, it is recommendable to consider it separately. as many RG meta-analyses have applied OLS methods without declaring the statistical model assumed.

An extension on the different statistical models that have frequently been assumed in this type of meta-analysis can be found in Chapter 2.

1.2.2.3 Other statistical methods

The previous sections have dealt with meta-analytical analysis from the conventional perspective, that is, the one that has been applied in the vast majority of the literature. Nevertheless, this perspective does not consider the dependency relationships that may arise between the scores.

When a test has a multidimensional structure with different subscales, all forming part of the same psychological construct and, in addition, these scales are applied to different groups within the same scientific study, dependence between scores can arise. This means that an observed effect size may be providing information about the direction or degree of deviation of another effect from that expected by the model (Assink & Wibbelink, 2016; Van den Noortgate et al., 2013). This dependence is not considered in the traditional meta-analysis approach to reliability generalization.

A solution to the dependency problem is to try to model it through multivariate models such as the one proposed by Raudenbush et al. (1988). Another way of modelling it is also through the application of multilevel models. These alternative models will be explained in more detail in chapter 4.

1.2.3 Replicability and Reproducibility of Reliability Generalization Meta-Analysis

Throughout this chapter it has been noted that there are multiple strategies available when carrying out a reliability generalization meta-analysis. Consequently, variability in the results may arise depending on the strategy followed. One of the advantages, given that this type of methodology does not have a pre-established analysis protocol, is that it allows for flexibility in the statistical analysis of the empirical data collected and the inferential conclusions to be drawn from the results. However, such flexibility implies that meta-analysts are aware of the importance of reporting the method of the work in a detailed and accurate form, specifying each of the decisions that have been taken to obtain these results. If the researcher is not transparent about the reporting, it makes the reproduction of the study complicated and, in some cases, impossible. It should not be forgotten that both the reproducibility and replicability of scientific research are among the fundamental pillars of the scientific method and are essential to the robustness of the conclusions.

As a consequence, and taking into account that to date no work has been published on this issue, in Chapter 2 we have conducted a study comparing the meta-analytical results obtained when applying different analysis strategies. Specifically, the techniques that are commonly used to carry out a meta-analysis of reliability generalization have been collected and their results have been compared both for the calculation of the average reliability coefficient and for the calculation of its confidence interval. Based on this work, we will be able to determine whether these different strategies can really influence the conclusions of the study. It should be noted that this study has been carried out on real data from published meta-analyses.

Following the thread of Chapter 2, we proposed to study the reproducibility indices in this type of meta-analysis. Especially in recent years, different authors (Artner et al., 2021; Hardwicke et al., 2018; Lakens et al., 2016; Maassen et al., 2020; Nosek et al., 2022) have published on the replicability and reproducibility of meta-analyses, mostly in the field of psychology. However, in the field of reliability generalization, no study or approach has been published on reproducibility rates in this specific type of meta-analysis. Thus, Chapter 3 is a reproducibility study of the reliability generalization meta-analyses included in Chapter 2.

Finally, Chapter 4 presents the last study that composes this doctoral thesis. Following the thematic on the incidence of the different analysis strategies when computing a reliability generalization meta-analysis, it seems relevant to know whether, when we are carrying out a meta-analytic study that implies, by its own nature, that dependence relationships may arise between the scores, the results obtained through an RG meta-analysis performed in the conventional method, that is, abolishing such dependence and separating the analyses, differ statistically from the results obtained through multilevel models that try to model that dependence and deal with it. In this study, different statistical models within the multilevel perspective have been compared with the conventional model of analysis, manipulating the conditions of the scales that have been included. This study has also been carried out on real empirical data from published meta-analyses, as the aim was not to identify which model works *best*, but to assess whether there are differences between different analytical decisions.

Chapter 2

Study 1:

“Reliability Generalization Meta-analysis: Comparing Different Statistical Methods”

2.1 Introduction

To date, a large number of RG meta-analyses have been carried out in psychology on different measurement instruments. A systematic search has identified more than 150 RG meta-analyses conducted on psychological measurement tools between 1998 and 2019 (Sánchez-Meca et al., 2019). Examples of RG meta-analyses are those of the *Beck Depression Inventory* (Yin & Fan, 2000), the *Childhood Autism Rating Scale* (Breidbord & Croudace, 2013), the *Yale-Brown Obsessive-Compulsive Scale* (López-Pina et al., 2015), or self-report measures of muscle dysmorphia (Rubio-Aparicio et al., 2020). As mentioned in the previous chapter, a reliability generalisation meta-analysis can be carried out using different analytical strategies. Thus, it is possible that by applying different strategies, different results may also arise.

2.1.1. Purpose

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

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

2.1.2. Statistical methods in RG meta-analysis

In this meta-review we have selected those transformations that have been most frequently found in RG meta-analyses. Table 1 presents the different methods to transform internal consistency coefficients with their corresponding sampling variances, as well as formulas to back-transform the transformed coefficients to the original metric (Sánchez-Meca et al., 2012). Note that in Table 1 the typical symbol to represent Cronbach's alpha reliability coefficients is used ($\hat{\alpha}_i$), as most RG meta-analyses use this coefficient to estimate the internal consistency of scales. This is due to alpha coefficients being routinely reported in primary studies. However, formulas shown in Table 1 can be applied to other types of internal consistency reliability coefficients.

Table 1.

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

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

Note: $\hat{\alpha}_i$: alpha coefficient reported in the i_{th} study. n_i: sample size of the i_{th} study. J_i: number of items of the test version used in the *i*th study. k: number of alpha coefficients of the RG meta-analysis. ¶The sampling variance formula for the untransformed internal consistency coefficients is that proposed by Bonett (2002). ln: natural logarithm. Hakstian-Whalen: Hakstian and Whalen's (1976) transformation. Bonett: Bonett's (2002) transformation.

The next decision, as previously mentioned, concerns the statistical model to be assumed. We have already seen the theoretical differences in each of the models presented. However, it is also well known that assuming one or another statistical model has consequences on how statistical analyses are accomplished and on the degree of generalizability of the meta-analytic results. In particular, how the reliability estimates are weighted is different depending on the statistical model assumed. Under an FE model the optimal weighting factor is the inverse of the sampling variance of each reliability coefficient, $w_{\mathrm{i}}^{\mathrm{FE}} = 1/V(y_i)$, with $V(y_i)$ being the within-study sampling variance of the reliability coefficient of the *i*th study. Alternatively, under an FE model the meta-analyst can decide not to weight the reliability estimates, that is, $w_{\mathrm{i}}^{\mathrm{FE}} = 1$. Under an RE model, the optimal weights are defined as the inverse of the sum of the sampling variance and the between-studies variance, $w_{\mathrm{i}}^{\mathrm{RE}} = 1/[V(y_i) + \tau^2]$, with τ^2 being an estimate of the between-studies variance (Borenstein, 2019; Cooper et al., 2019). Alternatively, an RE model can be applied by weighting the reliability coefficients by its sample size instead of its inverse variance (Schmidt & Hunter, 2015). On the other hand, regarding the VC model, as it is theoretically in a middle point between the FE model and the RE model, its results can only be generalized to a set of studies with identical characteristics to those of the studies included in the meta-analysis, and the optimal estimate of the average reliability coefficient implies not weighting the individual coefficients ($w_{\mathrm{i}}^{\mathrm{VC}} = 1$). The mathematical formulation of the three statistical models can be found in Table A2.1 in Appendix 2A.

Regardless of the statistical model assumed, in an RG meta-analysis it is usual to calculate an average reliability coefficient, its sampling variance, and a 95% confidence interval to estimate the average population reliability coefficient. Veroniki et al. (2019) have identified 15 alternative methods to construct a confidence interval for the average effect size under an RE model. Out of the large number of methods to construct confidence intervals, those usually applied in RG meta-analysis have been selected to be compared in this study. These methods are presented in Table 2. The methods differ on whether to transform the reliability estimates, the statistical model assumed, and how to weight reliability coefficients. Thus, we have considered methods under the FE, RE, and VC statistical models. In addition, we have included methods based on ordinary least squares (OLS). OLS method consists of applying conventional statistical methods, that is, to calculate an unweighted mean of reliability coefficients, to estimate its sampling variance, and to construct a 95% confidence interval as if the reliability estimates were single data from a sample of participants. Although the OLS method can be thought of as an FE model, here we consider it separately, as many RG meta-analyses have applied OLS methods without declaring the statistical model assumed. Note that in OLS and FE methods the reliability coefficients can be transformed or not to normalize their distribution and stabilize variances (in Table 2 the term 'y_i' interchangeably represents the transformed or untransformed reliability coefficient of the *i*th study). Under the VC model advocated by Bonett (2010), the average of the population reliability coefficients is estimated by calculating an unweighted average of the untransformed internal consistency coefficients; however, to construct a 95% confidence interval the average reliability coefficient must be transformed by Bonett's method. Table 2 also shows three methods under an RE model. The standard RE method implies estimating the sampling variance of the average reliability coefficient as the inverse of the sum of the weights (w_i^{RE}) and a standard normal distribution to construct a confidence interval (Konstantopoulos & Hedges, 2019). Following Schmidt and Hunter's (2015) approach, the REn method consists of not transforming the reliability coefficients and weighting them by the sample size of each study. Finally, the REi method is based on an improved method proposed by Hartung and Knapp (2001) to estimate the sampling variance of an average effect size and to assume a Student *t*-distribution with degrees of freedom equal to *k* – 1, *k* being the number of studies (Sánchez-Meca & Marín-Martínez, 2008). REi method offers better adjustment to the nominal confidence level than the RE and REn methods, as it takes into account uncertainty in estimation of between-studies variance, τ^2 (Hartung & Knapp, 2001; Rubio-Aparicio et al., 2018; Sánchez-Meca & Marín-Martínez, 2008; Veroniki et al., 2019).

Table 2.

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

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

Note: OLS: Ordinary Least Squares method. FE: Fixed-Effect model. VC: Varying-Coefficient model. RE: Random-Effects model. REi: Random-Effects model with the improved method of Hartung and Knapp (2001). REn: Random-Effects model weighting by sample size. y_i = transformed or untransformed reliability coefficient of the *i*th study. $\hat{\alpha}_i$ = untransformed internal consistency reliability coefficient of the *i*th study. n_i = sample size of the ith study. *k* = number of studies. S_y^2 = variance of the *k* transformed or untransformed reliability coefficients. $t_{k-1,\alpha/2}$ = (α/2)x100% percentile of the Student *t*-distribution with *k*-1 degrees of freedom. $z_{\alpha/2}$ = (α/2)x100% percentile of the standard normal distribution. $b = ln[\bar{n}/(\bar{n} - 1]$, *ln* being the natural logarithm and $\bar{n}$ being the harmonic mean of the sample sizes: $\bar{n} = k/\sum(\frac{1}{n_i})$.

Statistical theory predicts OLS methods as exhibiting the largest confidence widths, as they do not take advantage of cumulating the sample sizes of the primary studies when computing a confidence interval for the average reliability coefficient. They are followed by REi method, as it takes into account two sources of error among the reliability estimates (within- and between-study variability) and uncertainty in estimating the between-studies variance. RE and REn methods will offer narrower confidence widths than REi method, as they do not consider uncertainty in the estimation of the between-study variance. The VC method will present narrower confidence widths than the three RE methods, as it does not aim to estimate an average reliability coefficient from a super-population of potential reliability coefficients, but the average population coefficient of the studies included in the RG meta-analysis. Finally, the FE method will exhibit the narrowest confidence widths, as it assumes that the reliability estimates share a common population reliability coefficient (Sánchez-Meca et al., 2012).

2.2 Method

2.2.1. Study selection criteria

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

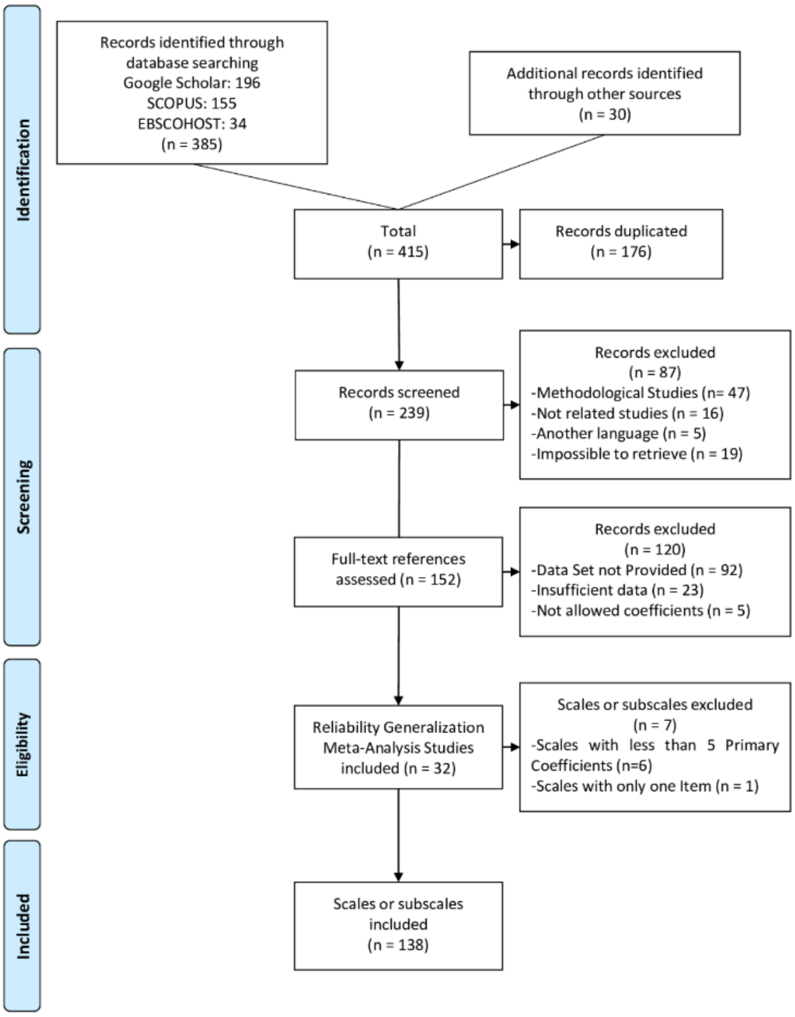
2.2.2. Search strategy

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

Figure 1 presents a flow diagram outlining the selection process of studies. The electronic searches yielded 385 references. Additional informal searches produced another 30 references. On discarding duplicated references, a total of 239 references were identified as potentially eligible for this research. From these, 207 references were excluded for not fulfilling some inclusion criteria (e.g., methodological studies which did not focus on internal consistency coefficients, did not present the whole dataset with the individual reliability coefficients, the dataset contained less than 5 reliability coefficients, or the psychological tool had only one item). Therefore, 32 RG meta-analyses were included in this research. The references of the 32 RG meta-analyses selected are openly available in Appendix 2C. As many of these studies included several psychological tests, or one psychological test with different subscales, we were able to obtain 138 datasets comprising scales or subscales contributing 4,350 internal consistency coefficients. Although our purpose was to include any type of internal consistency coefficients, all RG meta-analyses selected for this research used only Cronbach’s alpha reliability coefficients.

Figure 1.

Flow diagram of study selection process.



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

2.2.3. Data extraction

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

2.2.4. Data analysis

Statistical methods shown in Table 2 were applied on each of the 138 datasets of alpha coefficients. To estimate the between-studies variance (τ^2), DerSimonian and Laird's (DL) moments method was applied because it is one of the most widely used, although it is not the best one (cf. Blázquez-Rincón et al., 2023; Boedeker & Henson, 2020; Langan et al., 2017; Sánchez-Meca et al., 2012; Sánchez-Meca & Marín-Martínez, 2008; Veroniki et al., 2016; Viechtbauer, 2005). In order to assess whether the τ^2 estimator can affect the results of an RG meta-analysis, the restricted maximum likelihood (REML) estimator was also applied. Thus, sensitivity analyses were carried out by means of ANOVAs (one for the average alpha coefficient and the other for its confidence width) comparing the meta-analytic results for DL and REML τ^2 estimators. With the purpose of assessing whether the τ^2 estimator has an influence on the meta-analytic results, two-way ANOVAs with repeated measures in the two factors were applied, taking the average alpha coefficient and the confidence width as the dependent variables. The two factors being the τ^2 estimator (DL vs. REML) and the transformation method. Four transformation methods of the reliability coefficients were considered (not transformation, Fisher's Z, Hakstian and Whalen's and Bonett's transformations) and six statistical models: OLS, FE, VC, and three RE models (standard RE, REi, and REn models). Although a total of 24 combinations could be applied to obtain an average reliability coefficient, only 13 different methods were compared. This is due to the fact that VC (Bonett, 2010) and REn (Schmidt & Hunter, 2015) models do not admit coefficients to be transformed, therefore these statistical methods were applied for untransformed alpha coefficients only. In addition, note that RE and REi methods apply the same formula to calculate an average reliability coefficient (see Table 2). The difference between RE and REi methods is in how to construct a confidence interval. The

13 methods compared to obtain a combined reliability coefficient can be found in Table 2A.2 in Appendix 2A.

In addition, 18 different methods to calculate a confidence interval for the average reliability coefficient were applied (see Table 2). Out of these, 16 methods were obtained by combining the statistical models OLS, FE, RE, and REi with the four transformation methods (not transformed, Fisher's Z, Hakstian and Whalen's, and Bonett's transformations). Two additional methods were based on the VC model for Bonett's transformation and the REn model for untransformed coefficients. The 18 methods compared have been described in Table 2A.3 in Appendix 2A.

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

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

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

In order to compare the 13 alternative methods to calculate an average reliability coefficient and to compare the confidence width of the 18 methods to construct a confidence interval for the average reliability coefficient, two-way ANOVAs were applied. The two factors introduced in the model were the assumed statistical model and the transformation of the coefficients (with four conditions). For the average coefficient estimate, the two factors included in the ANOVA had four levels, while for the confidence width the statistical model had 6 conditions. In case of finding statistically significant results for any of the factors, post hoc comparisons were applied with Bonferroni's method.

The 13 methods to calculate an average reliability coefficient and the 18 methods to construct a confidence interval for the average reliability coefficient were computed in four different metric scales: those of the untransformed alpha coefficient and the three transformation methods (Fisher's *Z*, Hakstian and Whalen's and Bonett's transformation). In order to make them comparable, results for the three transformation methods were back-transformed to alpha metric by means of the formulas presented in Table 1.

The 138 meta-analytic datasets as well as the script codes used to analyse them are openly available at: https://bit.ly/vtgf7. All meta-analytic calculations were programmed in R (R Core Team, 2020). Shapiro-Wilk's normality test and skewness and kurtosis indices were calculated with the R package moments (Komsta & Nomovestky, 2015). ANOVAs and post hoc comparisons were carried out with the statistical programs IBM SPSS Statistics (v28; IBM Corp, 2021) and JAMOVI (v2.2; The Jamovi Project, 2021). Finally, to illustrate the results, multiple violin displays were constructed with the package ggplot2 in R (Wickham, 2016).

2.3 Results

2.3.1 Characteristics of the meta-analytic datasets

The 138 RG datasets were extracted from 32 studies that fulfilled our inclusion criteria. The RG datasets had a number of studies (*k*) that ranged between 5 and 319 primary studies or alpha coefficients, with an average of 31 primary studies (Median = 14 studies; Q_1 = 9; Q_3 = 319). The histogram of the number of studies showed a clear positive asymmetry, with 70.3% of datasets exhibiting fewer than 30 primary studies ($k < 30$) and only 6 datasets (4.3%) with *k* larger than 100. Sample sizes distribution of the more than 4,500 alpha coefficients ranged between 38 and 799, with a mean of 209 (Median = 220; Q_1 = 125; Q_3 = 249). A summary of the descriptive statistics for both number of studies and sample sizes can be found in Table 2A.4 in Appendix 2A. Figure 2A.1 is available also in Appendix 2A.

2.3.2 To transform or not to transform reliability coefficients

One controversial point in the RG meta-analytic arena is whether alpha coefficients should be transformed to normalize their distribution. To examine the extent to which different transformation methods achieved their objective of normalizing the alpha coefficient distribution, Shapiro-Wilk’s test and skewness and kurtosis statistics were calculated for each transformation method in each RG dataset. Table 3 presents the results. Regarding untransformed alpha coefficients, almost half of datasets (44.9%) reached statistical significance with Shapiro-Wilk's normality test, indicating a clear departure from the normality assumption. Compared to the untransformed alpha coefficients, the three transformation methods (Fisher's *Z*, Hakstian and Whalen's, and Bonett's transformations) substantially improved the normality adjustment of the alpha coefficient distribution, with rejection percentages of about 26%. In addition, the skewness indices for untransformed alpha coefficients (Table 3) clearly departed from symmetry (Mean = -.75; Median = -.71), whereas transformed coefficients improved the symmetry (Fisher's Z: Mean = .005, Media = .07; Hakstian and Whalen: Mean = .20, Median = .14; Bonett: Mean = -.09, Median = -.12). To determine whether these differences were statistically significant, a repeated-measures ANOVA was performed. The results confirmed these differences, $F(3, 411) = 31.1$, $p < .001$, $\eta^2 = .185$. Post hoc comparisons showed differences between no transformation of the coefficients and the three transformations, and between Hakstian and Whalen's and Bonett's transformation. Table 2A.5 in Appendix 2A presents the post hoc comparisons.

However, kurtosis indices for untransformed alpha coefficients were close to normality (Mean = 3.74, Median = 2.95), whereas those of the transformed coefficients led to slightly platykurtic distributions (Fisher's Z: Mean = 2.98, Median = 2.61; Hakstian and Whalen: Mean = 3.01, Median = 2.61; Bonett: Mean = 2.94, Median = 2.59). A repeated-measures ANOVA performed to compare the four transformation conditions yielded statistically significant differences, $F(3, 411) = 27.8$, $p < .001$, $\eta^2 = .169$, specifically between the coefficients without transforming and applying the three transformations. Table 2A.6 in Appendix 2A presents these results.

Table 3.

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

	S-W test	**Skewness**[¶]		**Kurtosis**[§]	
Transformation method	**Rejection percentage (*p* < .05)**	**Mean**	**Median**	**Mean**	**Median**
No transformation	44.9%	-.757	-.736	3.75	2.951
Fisher's Z	26.1%	-.017	.066	2.968	2.614
Hakstian-Whalen	26.8%	.19	.143	3.007	2.607
Bonett	26.8%	-.098	-.12	2.934	2.591

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

2.3.3 Between-study variance estimator

In order to examine whether the choice of the τ^2 estimator in an RE model could affect the average alpha coefficient and the confidence width, a sensitivity analysis was conducted consisting of applying two τ^2 estimators: DL and REML. This comparison only affected to the RE model for the average alpha coefficient and for the RE and REi models for the confidence width and the four transformation methods. The results can be found in Tables 2A.7-2A.13 in Appendix 2A. Regarding the average alpha coefficient, using DL or REML τ^2 estimators did not affect the results (see Table A8), $F(1, 137) = 1.11$, p = .294, $\eta^2 = .008$. However, an interaction between the τ^2 estimator and transformation method was found, $F(3, 411) = 26.29$, p < .001, $\eta^2 = .161$. Post hoc comparisons revealed statistically significant differences between the average alpha coefficient for DL and REML τ^2 estimators when alpha coefficients were not transformed (see Table A9). Regarding the confidence width, Table A10 presents the results as a function of the τ^2 estimator (DL vs. REML), transformation method, and statistical model (RE vs. REi). Table A11 presents the results of a three-way ANOVA. No statistically significant differences were found for the τ^2 estimator, F(1, 137) = 2.12, p = .147, η^2 = .015. Like with average alpha coefficient, a statistically significant interaction was found between the τ^2 estimator and transformation method, F(3, 411) = 4.50, p = .004, η^2 = .032, although with negligible proportion of variance accounted for. In fact, any of the post hoc comparisons for this interaction reached statistical significance (see Table A12). Similar results were found for the interaction between τ^2 estimator and statistical model (see Tables A11 and A13). As τ^2 estimator did not affect the results, meta-analytic calculations were presented using DL estimator only.

2.3.4 Averaging a set of reliability coefficients

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

Table 4.

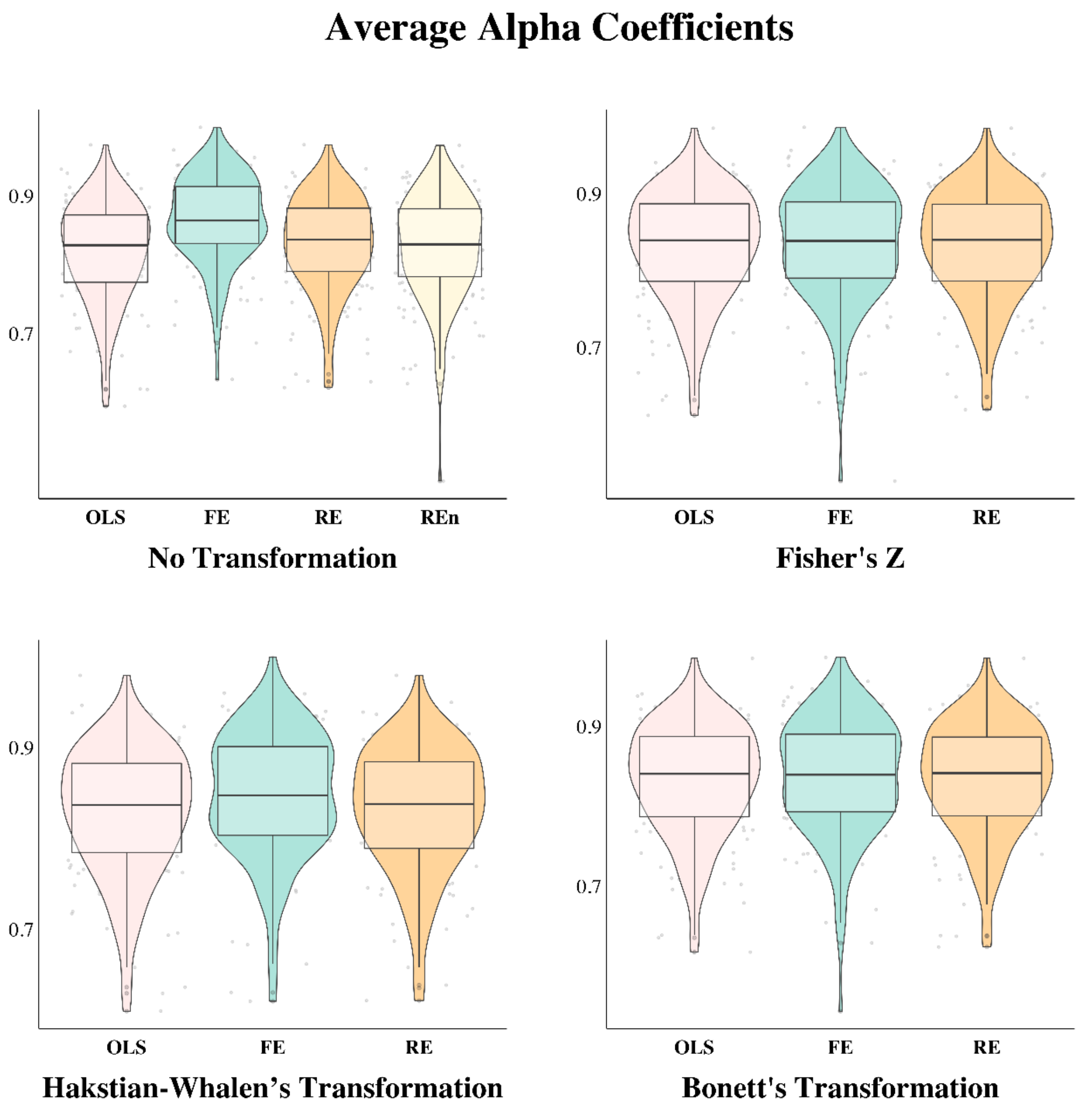
Results of average alpha coefficients for each analytic strategy

		Average Alpha							
Model	**Transformation**	**Mean**	***SD***	**Min**	***Q1***	**Median**	***Q3***	**Max**	**Range**
OLS	**No Transformation**	.819	.072	.595	.775	.829	.873	.974	.379
	Fisher's Z	.832	.07	.612	.786	.84	.887	.986	.373
	Hakstian-Whalen	.828	.07	.61	.785	.837	.883	.98	.369
	Bonett	.833	.069	.618	.787	.841	.888	.986	.368
FE	**No Transformation**	**.867**	.063	.634	.831	**.865**	.914	**1**	.366
	Fisher's Z	.836	.074	.527	.791	.839	.89	**.987**	.46
	Hakstian-Whalen	.848	.069	.621	.804	.848	.901	**1**	.378
	Bonett	.837	.072	.544	.793	.84	.891	**.987**	.443
RE/REi	**No Transformation**	.83	.07	.622	.791	.836	.883	.975	.353
	Fisher's Z	.833	.069	.62	.787	.84	.887	.986	.366
	Hakstian-Whalen	.832	.069	.622	.789	.838	.885	.98	.358
	Bonett	.834	.068	.624	.788	.842	.887	.986	.361
REn	**No Transformation**	.826	.076	**.486**	.783	.83	.881	.974	**.487**

OLS: Ordinary Least Squares model. FE: Fixed-Effect model. RE: Standard Random-Effects model. REi: Random-Effects model with the improved method of Hartung and Knapp (2001). REn: Random-Effects model weighting by sample size. Results for the VC model were not shown in this Table as they coincide with those of the OLS model with untransformed coefficients. Results for RE and REi models are coincident. *SD*: Standard Deviation. Min. and Max.: Minimum and Maximum average alpha coefficient. *Q1* and *Q3*: quartiles 1 and 3.

Figure 2.

Multiple violin and boxplots of the 13 different methods for averaging alpha coefficients.



Note: OLS = Ordinary Least-Squares model. FE = Fixed-Effect model. RE = Standard Random-Effects model weighting by the inverse variance. REn = Random-Effects model weighting by sample size.

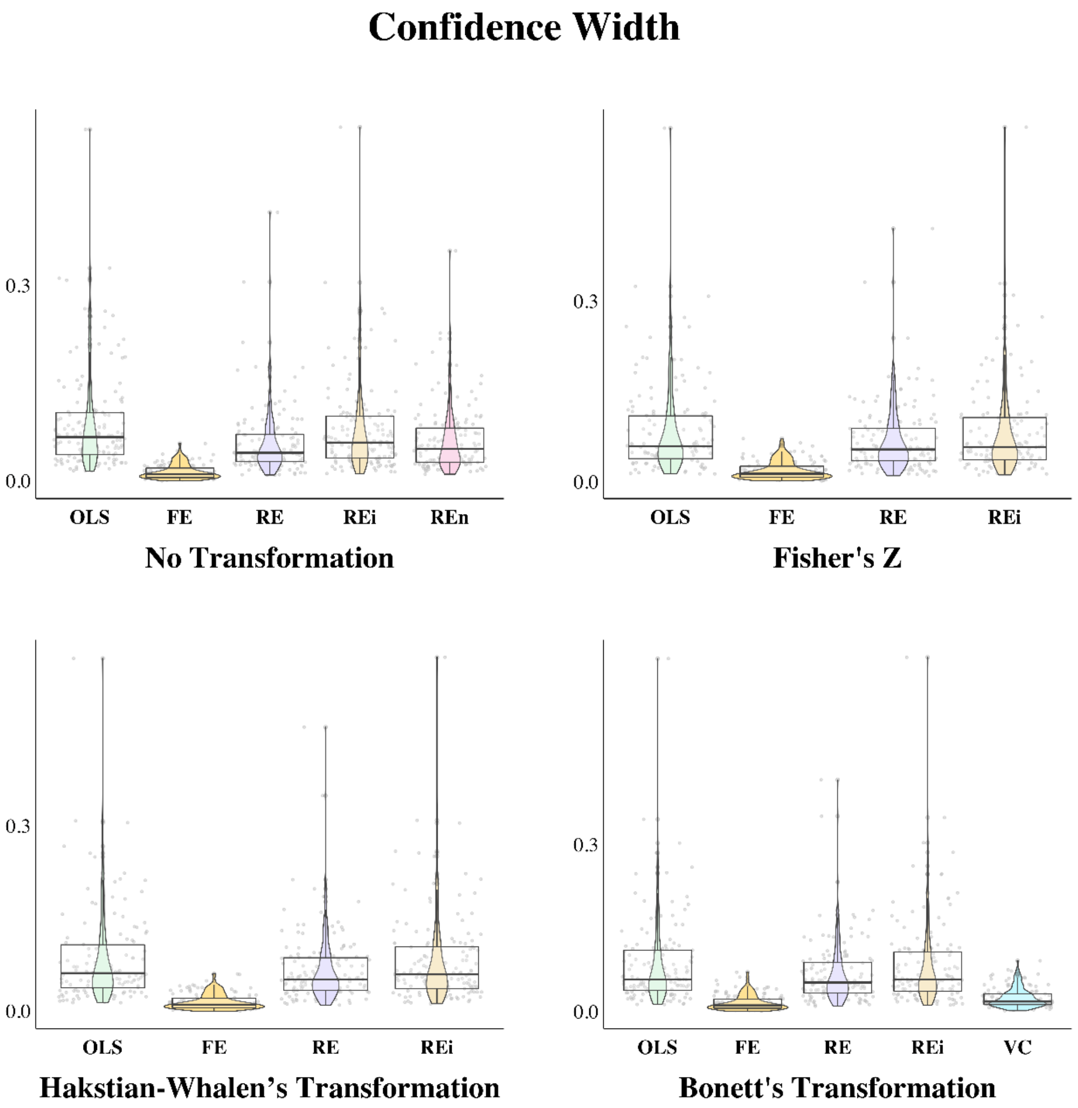
To compare methods among them, a two-way ANOVA was applied, with the average alpha coefficients as dependent variable and the statistical model and transformation method as factors. The results showed a statistically significant interaction between the two factors, $F(6, 1781) = 3.233$, $p = .004$, $\eta^2 = .011$, as well as the statistical model, $F(3, 1781) = 8.614$, $p < .001$, $\eta^2 = .014$. However, the proportion of variance accounted for by these factors was negligible (1.1% and 1.4%, respectively). Bonferroni's post-hoc comparisons indicated that significant differences were found between the FE model and the rest of the models (see Table 2A.14 in Appendix 2A). Specifically, significant differences were found between the untransformed average coefficients obtained assuming an FE model and the rest of the models, as well as within the FE model itself using Bonett’s and Fisher’s Z transformations (Table 2A.15 in Appendix 2A).

2.3.5 Constructing a confidence interval for the average reliability coefficient

Differences among the 18 methods to construct a confidence interval for the average reliability coefficient were also compared in terms of their confidence width. Table 5 presents descriptive statistics obtained by calculating the confidence width across the 18 analytical strategies. Both the mean and median indicated that larger confidence widths were found when OLS method was assumed without transforming the coefficients. While the lowest values were found under an FE model, the maximum values were found under OLS and REi models (i.e., RE model with the improved method of Hartung and Knapp). Figure 3 presents multiple violin and boxplots to illustrate the confidence widths through the different analytic methods compared. A two-way ANOVA was applied on the confidence widths as a function of the statistical model and transformation method. Statistically significant differences were found for the statistical model assumed, $F(5, 2466) = 108.675$, $p < .001$, $\eta^2 = .181$, but not for the interaction, $F(9, 2466) = .347$, $p = .959$, $\eta^2 = .001$, nor for the transformation method, $F(3, 2466) = .532$, $p = .66$, $\eta^2 = .00$). Regarding the multiple comparisons (see Table 2A.16 in Appendix 2A), a significant result appears between almost all models. Post hoc comparisons revealed statistically significant differences between all the different statistical models, with three exceptions only: FE vs. VC models, OLS vs. REi, and RE vs. REn.

Figure 3.

Multiple violin and boxplots of the 18 different methods for calculating the confidence width.



Note: OLS: Ordinary Least-Squares model. FE: Fixed-Effect model. RE: Standard Random-Effects model weighting by the inverse variance. REi: Random-Effects model with the improved method of Hartung and Knapp (2001). REn: Random-Effects model weighting by sample size. VC: Varying-Coefficient model.

Table 5.

Results of confidence widths for each analytic strategy.

		Confidence Width						
Model	**Transformation**	**Mean**	***SD***	**Min**	***Q1***	**Median**	***Q3***	**Max**
OLS	**No Transformation**	**.089**	.077	.015	.04	**.067**	.105	.54
	Fisher's Z	.085	.077	.013	.039	.059	.11	.587
	Hakstian-Whalen	.085	.075	.014	.038	.061	.108	.572
	Bonett	.085	.08	.013	.038	.058	.11	.635
FE	**No Transformation**	**.014**	.011	.000	.005	.01	.02	.057
	Fisher's Z	.019	.015	.002	.008	.013	.026	.072
	Hakstian-Whalen	.015	.012	.000	.006	.01	.021	.06
	Bonett	.016	.014	.001	.006	.011	.022	.071
RE	**No Transformation**	.059	.052	.009	.03	.043	.072	.412
	Fisher's Z	.07	.057	.01	.035	.054	.089	.421
	Hakstian-Whalen	.068	.059	.01	.034	.051	.086	.46
	Bonett	.069	.058	.01	.034	.052	.089	.417
REn	**No Transformation**	.062	.05	.01	.029	.049	.081	.353
REi	**No Transformation**	.079	.071	.011	.035	.059	.099	.543
	Fisher's Z	.084	.077	.012	.037	.058	.107	.589
	Hakstian-Whalen	.082	.075	.012	.036	.06	.104	.573
	Bonett	.084	.08	.012	.037	.058	.107	**.637**
VC	**Bonett**	.025	.018	.002	.013	.018	.032	.092

OLS: Ordinary Least Squares model. FE: Fixed-Effect model. RE: Standard Random-Effects model. REi: Random-Effects model with the improved method of Hartung and Knapp (2001). REn: Random-Effects model weighting by sample size. VC: Varying-Coefficient model. *SD*: Standard Deviation. Min. and Max.: Minimum and Maximum confidence widths. *Q1* and *Q3*: quartiles 1 and 3.

2.3.6 Assessing heterogeneity

To assess heterogeneity exhibited by a set of alpha coefficients, the I^2 index was calculated for each of the 138 RG datasets and for each transformation method, with the purpose of examining the extent to which different transformation methods lead to different I^2 indices. Table 6 and Figure 4 show the descriptive statistics of the I^2 indices and their distributions for each of the transformations. On average, I^2 index was over 90% in all transformation methods, except for Fisher's *Z* (88.21%). There was only one dataset with an I^2 value lower than 25% for Fisher's *Z* (I^2 = 14.64%). When this I^2 value was deleted from the analyses, the average I^2 for Fisher's *Z* slightly increased (from 88.21% to 88.75%) and its variability decreased (*SD* = 11.5 and 9.65, respectively). In the remaining datasets and transformation methods all I^2 indices exceeded 25%, and only a few showed I^2 values below 75%, the threshold usually established to assume high heterogeneity. Bonett's and Hakstian and Wallen's transformations performed very similarly. In addition, these two transformation methods yielded I^2 indices with lower variability (Range = 53.01% and 54.63%, respectively) than Fisher's Z and untransformed coefficients (Range = 84.77% and 60.32%, respectively).

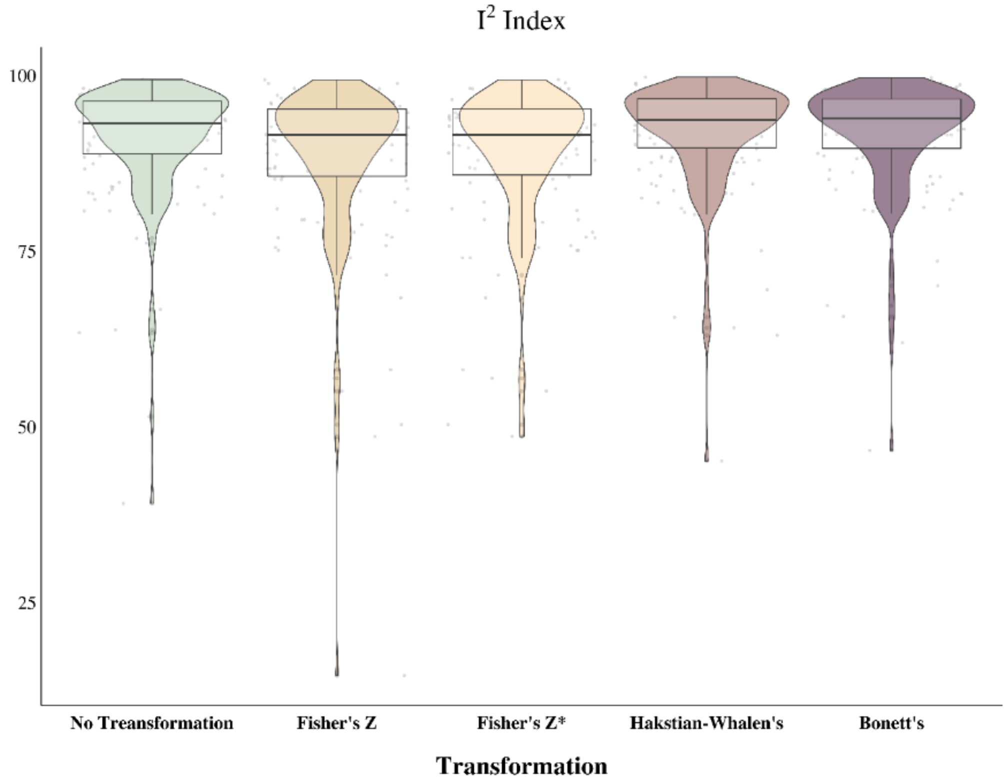
Table 6.
Results of aggregating the 138 I^2 indices for each transformation method.

	I^2 Index						
Transformation method	**Mean**	***SD***	**Min.**	***Q₁***	**Median**	***Q₃***	**Max.**
No Transformation	90.833	8.616	39.129	88.845	93.21	96.382	99.452
Fisher's Z	88.212	11.502	14.639	85.672	91.58	95.252	99.413
Hakstian-Whalen	91.7	7.826	45.174	89.741	93.68	96.723	99.797
Bonett	91.693	7.795	46.686	89.652	93.954	96.656	99.698
Fisher's Z¶	88.749	9.653	48.653	85.878	91.583	95.269	99.413

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

Figure 4.

Multiple boxplots of the I^2 indices for each transformation method.



Z* = Fisher’s Z once deleted the dataset with I^2 = 14.64%

To determine whether there were statistical differences in the I^2 indices as a function of the transformation method of the alpha coefficients, a repeated measures ANOVA was performed, finding statistically significant differences, $F(3, 411) = 66.6$, $p < .001$, $\eta^2 = .327$. Post hoc comparisons revealed statistically significant differences between all the transformation methods, with the exception of Hakstian and Whalen vs. Bonett transformations (see Table 2A.17 in Appendix 2A). Due to the presence of an outlier I^2 index (I^2 = 14.64%), another repeated measures ANOVA was also performed without it. However, deleting this outlier did not change the ANOVA results.

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

Figure 5.

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

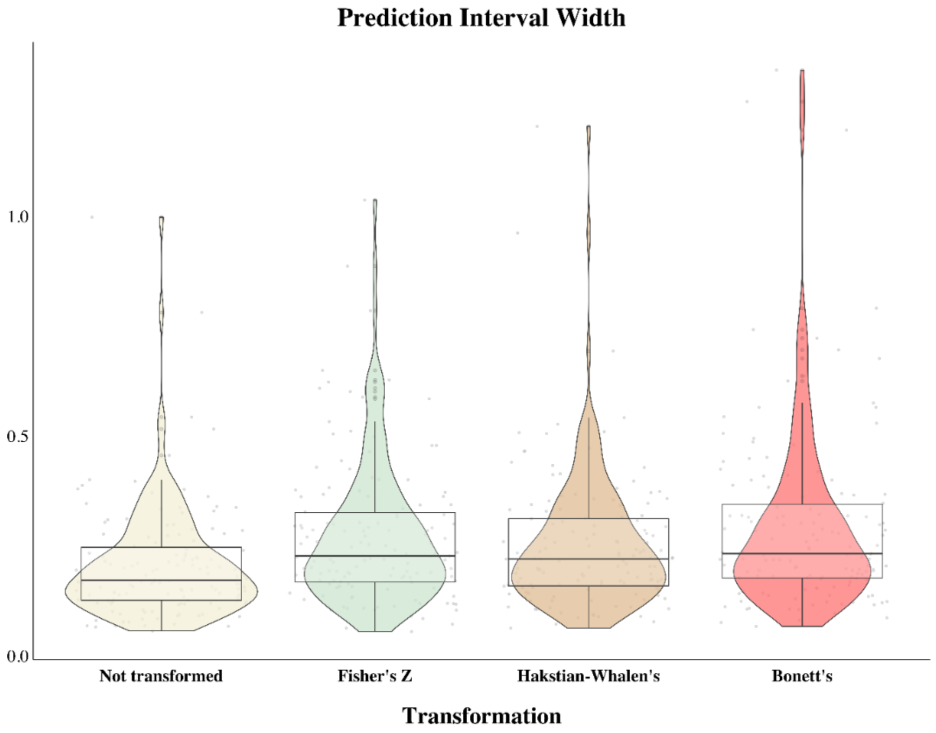


Table 7.

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

	95% Prediction Interval Width						
Transformation method	**Mean**	***SD***	**Min.**	***Q1***	**Median**	***Q3***	**Max.**
No Transformation	.207	.127	.057	.127	.172	.247	.998
Fisher's Z	.273	.163	.055	.168	.228	.326	1.037
Hakstian-Whalen	.254	.154	.063	.16	.221	.313	1.205
Bonett	.297	.209	.067	.177	.233	.345	1.333

Note: Hakstian-Whalen: Hakstian and Whalen's transformation. Bonett: Bonett's transformation. *SD*: Standard Deviation. Min. and Max.: Minimum and Maximum widths. *Q1* and *Q3*: quartiles 1 and 3.

To assess whether the transformation method of the alpha coefficients affected the width of prediction intervals, a repeated measures ANOVA was applied. The results showed statistically significant differences, $F(3, 411) = 43.2$, $p < .001$, $\eta^2 = .24$. Table 2A.18 in Appendix 2A shows the results of post-hoc comparisons, with statistically significant differences between all transformation methods. Larger interval widths were found with Bonett's transformation followed by Fisher's *Z* and Hakstian and Whalen's transformation.

2.4 Discussion

With the purpose of determining the extent to which different statistical methods used to integrate a set of reliability coefficients lead to different results, 138 datasets from 32 RG meta-analyses on psychological tests were analysed by applying multiple statistical methods developed in the meta-analytic arena. Regarding the different transformation methods of the reliability coefficients, our findings revealed that Fisher's Z, Hakstian and Whalen', and Bonett's transformations improved the normality adjustment of coefficient distribution than untransformed coefficients. Although the three transformation methods performed similarly, there are conceptual reasons for not using Fisher's Z to transform internal consistency coefficients like alpha and similar coefficients, as Fisher's Z was devised to transform correlation coefficients, whereas an internal consistency reliability coefficient is not a correlation coefficient, but a squared correlation coefficient (a ratio between true score and total score variance). Fisher's Z is adequate to transform test-retest reliability coefficients or parallel-forms coefficients, as these are calculated as correlation coefficients. For alpha coefficients, Hakstian and Whalen's and Bonett's transformations are most recommendable. Therefore, although there are proponents of not transforming reliability coefficients, transformation methods seem to normalize coefficient distribution, which is advisable as standard meta-analytic methods assume normality in their inference methods (cf., e.g., Borenstein & Hedges, 2019; Cooper et al., 2019).

RG meta-analyses always report an average reliability coefficient. Thirteen methods to calculate an average alpha coefficient were compared, depending on the statistical model assumed, weighting factor, and transformation method. An ANOVA applied with the statistical model and transformation of the coefficients as factors showed a statistically significant result for the interaction between them, as well as for the statistical model. However, the proportion of variance accounted for by these factors was negligible (about 1% only), revealing a limited influence. Post hoc comparisons indicated that the average alpha coefficients under an FE model were larger than those of other models (Table 4). REn model gave the lowest average alpha coefficients as well as the largest ones (from .487 to .974), exhibiting the largest variability. REn method consists of weighting the untransformed reliability coefficients by sample size. If in an RG meta-analysis reliability coefficients and sample sizes are correlated, then models that include sample size in the weighting factor can offer biased estimates of the average alpha coefficient. Our findings do not enable determine the extent to which different statistical models can lead to biased estimates of the population alpha coefficient in presence of alpha-sample size correlation, but an important recommendation when conducting an RG meta-analysis is to calculate the correlation between alphas and sample sizes. If a negative correlation is found, then we can assume that this RG meta-analysis can be suffering what in the meta-analytic arena is usually named ‘small study effects’, that is, studies with small sample sizes present higher alpha coefficients than larger ones (Rothstein et al., 2005). Through the 138 RG datasets, correlations between alpha coefficients and sample sizes varied from -.79 to .73, with Median equal to .06 (Mean = .07, *SD* = .28). These results evidenced that it is usual to find positive or negative correlations between alphas and sample sizes in RG meta-analyses. Therefore, in presence of correlation it is very advisable to apply methods to assess whether ‘small study effects’, ‘reporting bias’, ‘publication bias’ or any other biasing factors can affect the meta-analytic results. Such techniques as funnel plots, Egger’s test, or trim-and-fill methods should be applied to assess whether these biasing effects can affect the meta-analytic results (Rothstein et al., 2005; Vevea et al., 2019). In case of a negative correlation between alphas and sample sizes, the meta-analyst can decide to apply the FE weighting factor to calculate an average alpha coefficient, as this model is less likely to give biased estimates.

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

A confidence interval for the average reliability coefficient is also typically reported in RG meta-analyses. A total of 18 alternative methods to construct confidence interval for the average alpha coefficient were compared, in terms of the confidence width. Coinciding with previous research, our findings indicated that the different transformation methods of the alpha coefficients barely affected the confidence width for a given statistical model (Romano et al., 2010). However, the statistical model assumed dramatically affected confidence width. ANOVA results showed statistically significant differences as a function of the statistical model, with a proportion of variance accounted for of medium to large magnitude ($\eta^2 = .181$).

The largest confidence widths were obtained with the OLS methods, as they do not take advantage of the accumulation of sample sizes through the studies. On average, REi method was that which exhibited confidence widths more similar to those of the OLS method. As expected, the RE and REn methods on average exhibited narrower confidence widths than the REi and OLS methods, with average confidence widths varying between .059 and .070. Unlike the REi method, the RE and REn methods do not consider the uncertainty in estimating the between-studies variance, providing narrower confidence intervals than those of REi method (Sánchez-Meca & Marín-Martínez, 2008; Sidik & Jonkman, 2002; Stijnen et al., 2021). Both the FE and VC models exhibited the narrowest confidence intervals. The confidence width of VC model was, on average, .025, whereas under the FE model the average confidence widths varied between .014 and .019, being the narrowest widths of all models. The reasons for such narrow confidence widths are different for VC and FE methods. The FE model considers that all studies are estimating a common population reliability coefficient implying that the statistical calculations only take into account one error source: that due to sampling of participants (Borenstein et al., 2009; Sánchez-Meca et al., 2012). The VC model obtains narrower confidence widths than OLS and RE models as this model does not assume that the reliability coefficients from the studies are one random sample of a larger super-population of potential reliability coefficients (Bonett, 2010).

Parameters under the RE model can be estimated by means of alternative estimators. In particular, a large number of between-study variance estimators have been proposed (cf., e.g., Blázquez-Rincón et al., 2023). Comparing the results of applying different variance estimators was beyond the scope of this study. Nevertheless, we compared the results for two between-study variance estimators, DL and REML estimators, which are the most commonly used. Negligible differences were found on the calculation of the average alpha coefficient.

Regarding variability of reliability coefficients, I^2 indices revealed large heterogeneity in most RG datasets, indicating that reliability estimates reported in primary studies are affected by such study characteristics as composition and variability of samples and methods and context of application. In addition, heterogeneity was maintained regardless of the transformation method of the alpha coefficients. Therefore, the search for study characteristics that can explain heterogeneity is warranted in practically any RG meta-analysis. An additional finding was found about prediction intervals under an RE model. The width of the prediction intervals clearly varied as a function of the transformation method of the alpha coefficients, with wider intervals when Bonett's transformation was applied, followed by Fisher's Z and Hakstian and Whalen's transformation. Therefore, the choice of the transformation method is an important decision to interpret the width of the prediction intervals in an RE model.

2.4.1 How to select the statistical model?

If, as our findings evidence, the selection of the statistical model greatly affects the meta-analytic results, then an important question concerns the arguments that must guide the selection of the statistical model. It is important to note that our investigation does not enable determining which statistical model is most appropriate in an RG meta-analysis, as we have not conducted simulation studies, but empirical research based on real RG datasets. Therefore, our recommendations in this section are not based on our findings, but on previous theoretical work and results of simulation studies. The main question which must guide selection of statistical methods in an RG meta-analysis is to what extent the meta-analyst intends to generalize their results as well as the heterogeneity exhibited by the reliability coefficients. If the aim is to generalize to a set of studies with identical characteristics to those of studies in the meta-analysis, then the FE or the VC models are most recommendable. To decide between FE and VC models, the key question is whether the reliability estimates obtained in the primary studies exhibit heterogeneity. If this is not the case, then the FE model is most appropriate. However, if the reliability estimates exhibit heterogeneity among them, then VC should be chosen. How can we determine whether a set of reliability coefficients are heterogeneous? Several methods can be applied, such as the calculation of the I^2 index, such that if I^2 is larger than 25%, there is evidence of heterogeneity. Another method consists of testing the homogeneity hypothesis with Cochran's Q statistic, such that if the Q statistic reaches statistical significance (e.g., $p < .05$) there is evidence of heterogeneity. Other related methods involve calculating a prediction interval around the average reliability coefficient, or interpreting the magnitude of the between-studies standard deviation, τ (Borenstein, 2019; Stijnen et al., 2021). Our results evidenced that RG meta-analyses exhibit large heterogeneity (I^2 indices clearly over 25% and prediction intervals were wider than confidence intervals). As a consequence, FE models will be warranted in exceptional cases only. Even in the presence of apparent homogeneity, applying this model will be risky because heterogeneity statistics may have low power when the number of studies is small. Regarding OLS methods, we included it in our comparisons because they have been applied in many RG meta-analyses published in psychology. However, their application in RG meta-analysis, like in other kinds of meta-analysis, is not advisable under any circumstances, as they do not take into account the distributional properties of the reliability coefficients, leading to misspecification errors. RG meta-analyses that have estimated their parameters using OLS may have achieved clearly different results than if they had applied RE, VC or FE models.

When the meta-analyst intends to generalize their results to a larger population of studies with similar but not exactly identical characteristics to those of the studies included in the meta-analysis, then an RE model is best. From the three RE models here described, the RE, REi, and REn models, the REi model should be mainly chosen. This is because this model takes into account the uncertainty in estimating the between-studies variance (τ^2). However, to be adequately applied, RE models need several assumptions to be fulfilled: normality of the true reliability coefficient distribution, a stable estimate of the between-studies variance, and random sampling of studies from a larger population of primary studies. Strictly speaking, random sampling assumption cannot be met, as studies included in an RG meta-analysis are never randomly selected from a larger population of potential studies. Nevertheless, it is sufficient if the meta-analyst can reasonably assume, under a conceptual basis, that studies included in an RG meta-analysis are a representative sample of the super-population of primary studies; for example, when there is not correlation between alpha coefficients and sample sizes, or there is not publication bias, small study effects, nor other potential biasing factors (Laird &

Mosteller, 1990; Sánchez-Meca et al., 2012). On the other hand, the normality assumption can be relaxed, as recent simulation studies have demonstrated that RE and REi methods are not very affected by departures from normality (Kontopantelis & Reeves, 2012; Rubio-Aparicio et al., 2018). A more serious problem is to obtain an accurate estimate of the between-studies variance (τ^2). A meta-analysis with a small number of studies will have difficulty in accurately estimating τ^2. Note that τ^2 is an important parameter in calculating an average reliability coefficient and to construct confidence intervals and prediction intervals around it. To warrant a stable estimate of τ^2, results from previous simulation studies recommend applying RE and REi methods for meta-analyses with more than 20 studies (Aguinis et al., 2011; Sánchez-Meca et al., 2012). RG meta-analyses with fewer than 20 studies and in the presence of heterogeneity should apply REn method, as it is not necessary to estimate τ^2, provided reliability coefficients and sample sizes are not correlated. Otherwise, the VC model should be the most reasonable choice and the meta-analyst should limit results generalization to studies included in the meta-analysis only.

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

2.4.2 Limitations of study

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

Finally, the main limitation of our investigation was that our findings were not based on the results of a simulation study, but on empirically comparing meta-analytic results from real databases. We devised our study as a previous step to carry out future simulation studies comparing the performance of the different statistical methods to address the typical results in an RG meta-analysis. Our results can be useful for future simulation studies in two ways. First, it was important to know whether different analytic methods applied to real RG meta-analyses exhibit relevant differences in the meta-analytic results (in terms of average reliability coefficient, confidence interval, heterogeneity, and so on). If different statistical methods to synthesize reliability coefficients exhibit only negligible discrepancies, then to carry out a simulation study might not offer useful answers. Second, our results can help researchers interested in carrying out future simulation studies in to design the manipulated conditions based on real characteristics of RG meta-analyses typically published in psychology (e.g., in terms of number of reliability coefficients, average reliability, sample sizes of the single studies, heterogeneity variance, etc.). Thus, future simulation studies can base their parameter conditions on our findings. Descriptive statistics reported in tables in the paper as well as in the Supplementary file will be useful for this purpose.

2.4.3 Future research

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

2.5 Conclusion

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

Chapter 3

Study 2:

“Reliability Generalization Meta-*Analysis: A Reproducibility Study”*

3.1 Introduction

The proliferation of empirical evidence and scientific studies must be supported by analytical techniques and strategies that make it possible to synthesize results and conclusions that can be generalized. As we have already seen, meta-analysis has emerged as a tool for compiling and synthesising information from a multitude of empirical studies on the same subject in order to extract generalisable results. One of the distinctive features of meta-analysis is that it does not have a single analytical strategy, but rather there are numerous statistical techniques for conducting it that depend on and vary according to the primary data collected and the final objective of the study.

Due to the number of decisions that the meta-analyst has to make, it is important that all of them are carefully considered and taken, and always reported with total clarity and transparency. This implies, in addition to verifying the use of good methodological practices, ensuring that the work is reproducible and opened to the scientific community (Maassen et al., 2020).

Discussing reproducibility implies a proper distinction between the terms replicability and reproducibility. *Reproducibility* is based on reusing the same data and strategies as the original researcher to check that the results are consistent, while *replicability* is about starting the whole process again, beginning with the initial search, and verifying whether the results are congruent with those obtained in the first study (Artner et al., 2021). Focusing on the specific type of meta-analysis under study in this thesis (Reliability Generalization Meta-Analysis -RG MA-), an RG reproducibility study would imply redoing all the analyses using the database of the original study following the same procedure as stated by the investigator. A replicability study, in contrast, would involve starting the whole process over again. In a replicability study it is not necessary to follow each step of the method as specified by the original researcher, however it is essential to contrast whether the results obtained are congruent or not with the original ones, being a successful replicability if both strategies result in congruent values. It is important to emphasize that one of the fundamental principles of the scientific method is the possibility of replicating the experimental results, thus consolidating the conclusions drawn from them (Artner et al., 2021).

To date, no work has researched the reproducibility/replicability of RG meta-analyses. Nevertheless, Maassen et al. (2020) did investigate the reproducibility of meta-analyses within the field of psychology. The main result was that the reproducibility rate for primary effects was 55%, while the rate for meta-analyses was 61%. They also reviewed the reasons for this lack of reproducibility in both the reanalysis of the primary effects and the meta-analysis results and found that the main problems were the lack of information to carry out the reanalysis, the ambiguity of the report on the applied estimation methods and the lack of transparency on the extraction of the relevant effects chosen. However, regarding the irreproducibility rate of the meta-analyses, they also concluded that most of the discrepancies they collected were negligible. On the other hand, Artner et al. (2021) reproduced 70% of the primary claims (PC) of unpublished raw data from experimental studies in the field of psychology. They defined a PC as an a priori hypothesis that is evaluated through the null hypothesis significance test (NHST).

What is clear from all this is that for replication to be successful, it is imperative that the researcher is transparent about the whole research process. In fact, simply by maximizing transparency and reporting, the rate of replicability improves. And, if replicability increases, so does the robustness of the results and the credibility of the conclusions. Consistently maintaining a low replicability rate is a symptom of poor research practices and weak conclusions (Nosek et al., 2022).

3.1.1 Current Study

The main purpose of this study was to assess the extent to which the results of reliability generalization meta-analyses are reproducible. As a consequence, the transparency of reporting information on the reproducibility of the analyses was also evaluated. To achieve this, we collected RG meta-analyses published on psychological scales and constructs from 1998 to December 2020, which included openly, in the article itself, in a supplementary file or in an online repository, the database with all the necessary information extracted from the primary studies.

The meta-analytic strategies explained by the authors in the original meta-analyses were reproduced and two measures of comparison of the reproduced results were established: Pearson’s correlation between the reported values and the reproduced values, and a discrepancy index that determined the percentage by which the results differed from each other. By means of this index, three categories of reproduction success were determined: successful, approximate, and erroneous reproduction.

The variables to be reproduced were the average reliability coefficient and its confidence interval, and the two typical measures of heterogeneity in this type of work: I^2 index and Cochran's Q statistic.

3.2 Method

This study is an extension of the one carried out in the previous chapter, gathering the information we obtained from the databases collected and examining the reproducibility rate we found in this particular type of meta-analysis. All the materials necessary for the reproduction of this work are available in https://bit.ly/65w8nn

3.2.1 Study selection criteria and search strategy

As in the previous chapter, to be part of this research, meta-analytic studies had to fulfil several conditions: (a) the study had to be a meta-analysis of reliability generalization on one or more psychological scales; (b) the study had to provide, in the article itself, in an online repository or in a supplementary file, the complete database with the individual reliability coefficients, and their sample size, extracted from the primary studies; (c) the reliability coefficients analysed had to be Cronbach's alpha coefficients; and, finally, (d) all the articles had to be written in English or Spanish. In addition, and in order to carry out the analyses, at least two or more coefficients per scale or subscale had to be provided.

The electronic search was performed in the previous study and there was no modification for this study.

3.2.2 Data extraction

Data extraction was done following two independent strategies. First, data were extracted for the primary studies that comprised each of the meta-analyses. In this first step, the individual reliability coefficients, the sample size of each primary study and the number of items composing each scale/subscale were extracted from the databases provided by the meta-analyses.

On the other side, data were extracted on the method and results of these meta-analyses, essential to evaluate the transparency of the report, to establish the type of analysis used in each case and to check whether replication had been successfully achieved. In this case, we extracted the average reliability coefficients together with their confidence intervals (also the estimation method applied), the typical heterogeneity indices in this type of study (I^2 and Q), the statistical model assumed, the transformation of the coefficients applied if specified, and the statistical software in which the analyses were carried out.

In the case that a study performed several independent meta-analyses because it had different scales or the same test had different subscales, each of them was taken independently and analysed separately.

Data extraction was done with a single coder, at two different time moments. The results were checked for intra-rater agreement. In the event that the computed results differed from the reported results, the databases were rechecked to verify that the inconsistency was not due to problems in the coding.

3.2.3 Data analysis

Data reanalysis was carried out with the *metafor* package (Viechtbauer, 2010). The graphical part was performed with the packages *ggplot2* (Wickham, 2016), *patchwork* (Pedersen, 2022), *cowplot* (Wilke, 2020), and *forcats* (Wickham & RStudio, 2023). All these packages have been used in R software (R Core Team, 2020).

Each of the meta-analyses reproduced was analyzed following the process specified by the original authors in terms of the model assumed for the analysis, the model estimator used if random effects were assumed, and the transformation applied to the coefficients. In the case that other details of the analysis were reported, such as the weighting of the model or the method of estimating the confidence intervals of the average reliability coefficient, this was also considered. When the authors did not report information on any of these categories, the analyses were computed as follows: if they did not specify the model, the OLS method was assumed as the statistical model. If they reported that the model was a random effects model, but the heterogeneity variance estimator used was not specified, the DerSimonian-Laird (DL) estimator was computed, since it is the most widely used estimator, although it is not the most correct one (Langan et al., 2017; Sánchez-Meca et al., 2012; Sánchez-Meca & Marín-Martínez, 2008; Viechtbauer, 2005). And finally, if no transformation of the coefficients applied was reported, it was assumed that none had been applied and the raw coefficients were taken.

Having performed all the reanalyses, to determine the extent to which the results were reproductions of the original results, two testing strategies were carried out. First, the Pearson correlation between the reported values and the reproduced values was computed. In this way we could study the similarity between the reported and reproduced values. According to Cohen (Cohen, 1977), when the value of r_{xy} is greater than .5, the correlation is moderate, and when it exceeds .8, the correlation is significantly strong.

Furthermore, a discrepancy index has been calculated (Artner et al., 2021; Sánchez-Meca et al., 2012) that provided information on the percentage, in absolute value, of change between the reported value and the reproduced value, following the formula:

$$DI_j = \left|\left(\frac{\bar{Y}_{reproduced} - \bar{Y}_{reported}}{\bar{Y}_{reported}}\right)\right| \times 100, \tag{1}$$

where $\bar{Y}_{reproduced}$ refers to the value calculated here and $\bar{Y}_{reported}$ is the value reported in the original article. The results have been classified into three categories: *DI* < 5%, reproduced result; between 5 and 10%, approximate result; and, greater than 10%, error in reproduction, that is, the reproduced result is significantly discrepant from the reported value.

3.3 Results

3.3.1 Characteristics of the Meta-Analytic Studies

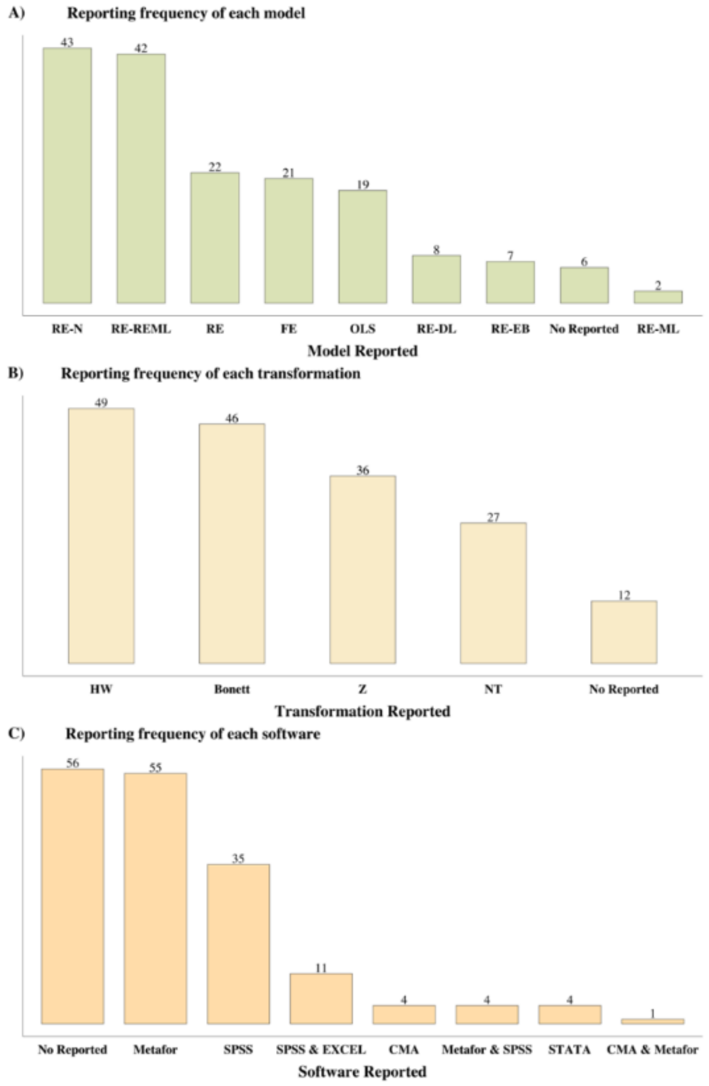
In the previous chapter, Figure 1 showed the flowchart with the study selection process. Although 152 studies initially met the criteria to form part of this study, 92 (60.53%) did not provide the database used to calculate the meta-analysis with primary study information, 23 (15.13%) did not provide sufficient information to reproduce these analyses and 5 (3.29%) did not employ the coefficients allowed in this study.

Finally, we had 32 studies that met all the inclusion criteria. From these 32 studies, we were able to extract 170 databases corresponding to independent meta-analyses of different scales or subscales. Regarding their characteristics, Figure 1 shows the statistical models, transformations and statistical software most frequently reported. The most frequently assumed model was the random-effects model weighted by sample size (RE-

N; 43 meta-analyses) and the random-effects model that used the restricted maximum likelihood estimator as the heterogeneity estimator (RE-REML; 42 meta-analyses). The most frequently used transformation was the Hakstian-Whalen transformation (49) and, finally, the most frequently used software was *metafor*. It should be noted that not reporting the software was more frequent (56) than the use of *metafor* (55).

Figure 1.

Reporting frequency of each model (A), transformation (B) and software (C)



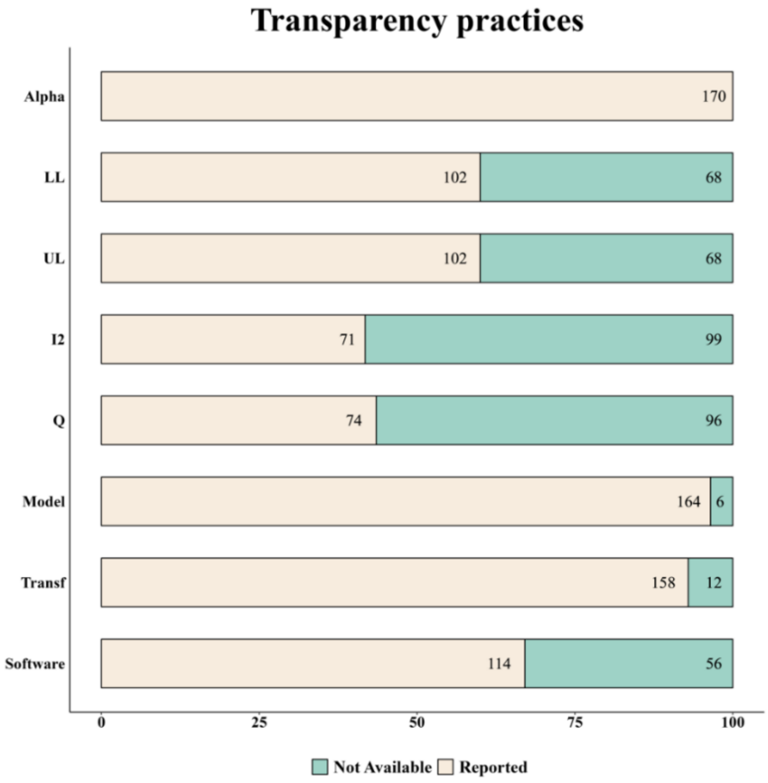
Note. RE-N = Random Effects model weighted by sample size. RE-REML = Random Effects model fitted with Restricted Maximum-Likelihood estimator. FE = Fixed-Effect model. RE = Random Effects model. OLS = Ordinary Least Squares model. RE-DL = Random Effects model fitted with DerSimonian-Laird estimator. RE-EB = Random Effects model fitted with Empirical Bayes estimator. RE-ML = Random Effects model fitted with Maximum-Likelihood estimator. HW = Hakstian and Whalen's transformation. Bonett = Bonett's transformation. Z = Fisher's Z transformation. NT = untransformed alpha coefficients. CMA = Comprehensive Meta-Analysis

3.3.2 Transparency Practices

Of the 170 scales/subscales that reported the average reliability coefficient computed, only 102 reported the confidence intervals of the average coefficient (60%). Regarding the heterogeneity indices, 71 provided the value of the I^2 index (41.76%) and 74 the value of the Q statistic (43.53%). Practically all the studies reported the statistical model assumed (164 meta-analyses; 96.47%) and the coefficient transformation applied (158; 92.94%). Finally, 114 meta-analyses (67.06%) reported the software used to compute the statistical analyses. In Figure 2 we can observe graphically the transparency rates for each category, the x-axis being the percentage value of the report and the inner value the absolute number of cases.

Figure 2.

Assessment of transparency practices for each variable.



Note. The x-axis shows the percentage of the total number of scales/subscales (N = 170). The internal values show the absolute values for each variable. LL = Lower Limit of the confidence interval. UL = Upper Limit of the confidence interval. Transf. = transformation of the coefficients.

3.3.3 Reproducibility Results

Although 170 scales/subscales were those that provided at least the average reliability coefficient, not all of them could be part of the reproducibility study: six of these subscales only provided one or two primary coefficients, making meta-analytic calculations impossible. In addition, the database provided by another study did not include data for one of the subscales reported in the article. That is, only 163 (95.88%) average coefficients, 95 (55.88%) confidence intervals, 68 I^2 indices (40%) and 71 Q statistics (41.76%) could be reproduced.

Table 1 shows the results of the descriptive statistics (mean, standard deviation, median, minimum and maximum values, and quartiles 1 and 3) for each variable both in its original form (*reported*) and in its recalculated form (*reproduced*). It is worth noting that both heterogeneity indices showed slight changes between the reported and reproduced. Specifically, for the mean value, in the case of the Q statistic ($Q_{Reported}$= 1076.04, $Q_{Reproduced}$= 1241.81), for the *SD* in the case of the I^2 index ($I^2_{Reported}$= 13.79, $I^2_{Reproduced}$= 7.48), and for the minimum values, in the case of both heterogeneity indices ($I^2_{Reported}$= 0, $I^2_{Reproduced}$= 56.17; $Q_{Reported}$= 0, $Q_{Reproduced}$= 6.84). The results for the rest of the variables are quite similar.

Table 1.

Descriptive Statistics for reported and reproduced results in each variable

		Mean	**SD**	**Min**	**Q1**	**Median**	**Q3**	**Max**
Reported	**Alpha**	.83	.07	.51	.79	.84	.89	.98
	CI$_{LL}$	.77	.12	.23	.73	.8	.84	.93
	CI$_{UL}$	.84	.09	.5	.81	.86	.91	.99
	***I*²**	90.05	13.79	0	91.39	94.1	95.94	98.31
	Q	1076.04	2250.7	0	123.19	482.72	927.83	12947.87

	Alpha	.83	.07	.5	.79	.84	.88	.97
	CI$_{LL}$	.78	.12	.23	.73	.8	.85	.93
Reproduced	**CI$_{UL}$**	.84	.1	.42	.81	.86	.91	.95
	***I*^{2}**	92.37	7.48	56.17	91.68	94.73	96.56	98.86
	Q	1241.81	2326.83	6.84	139.26	486.58	1151.32	12945.37

Note. CI: confidence interval. LL: lower limit. UL: upper limit. SD: Standard Deviation. Min. and Max.: Minimum and Maximum. Q1 and Q3: quartiles 1 and 3.

As explained above, two analysis strategies were carried out to determine the degree to which reproduction was successful. The descriptive results of the reproduction as a function of the discrepancy index are shown in Table 2. While the median remains below 5% discrepancy in all cases, the mean is above 10% in the case of the *Q* statistic (18.485). This is mainly due to the fact that the mean is a statistic that is very sensitive to the presence of extreme values, as we can see in the column of maximum values found, where this index presents the highest value (262.088), a value notably higher than the rest. However, the rest of the variables also obtained maximum values well above 10%, the lowest value being that of the I^2 index of heterogeneity (29.636).

Table 2.

Descriptive statistics of Discrepancy Index

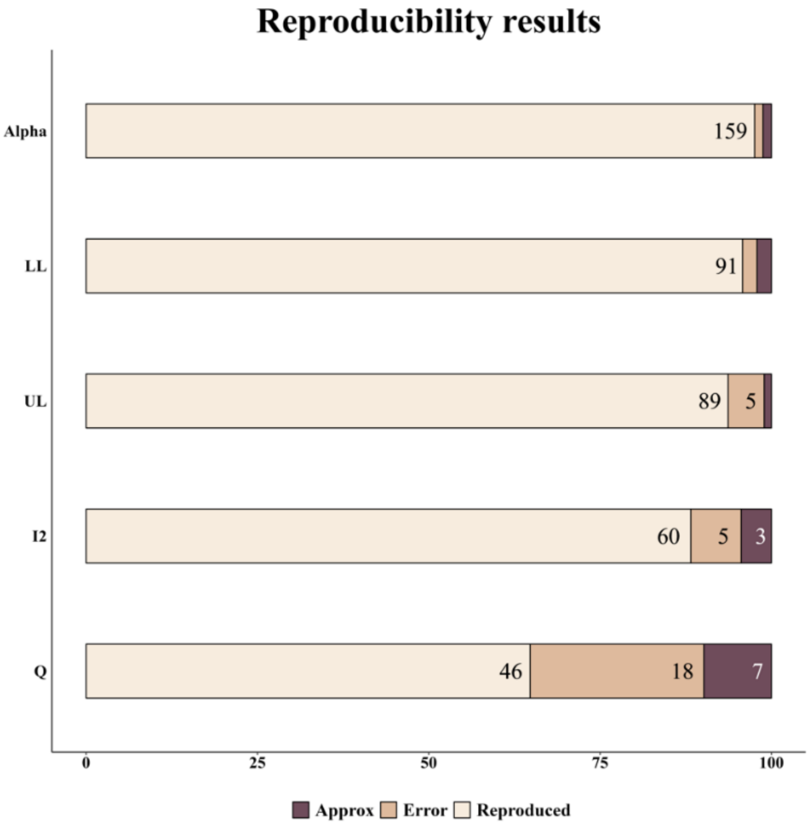
Variables	**N**	**Mean**	**SD**	**Min.**	**Q1**	**Median**	**Q3**	**Max.**	**N$_{>5\%}$**	**N$_{>10\%}$**
Alpha	163	1.394	5.774	0	.18	.443	.957	62.313	4	2
CI$_{LL}$	95	2.251	9.265	.004	.17	.463	1.07	79.077	4	2
CI$_{UL}$	95	2.143	6.935	.006	.179	.46	1.024	54.087	6	5
***I*^{2}**	68	2.509	6.207	0	.004	.093	1.219	29.636	8	5
Q	71	18.485	46.014	0	.002	.163	12.208	262.088	25	18

Note. CI$_{LL}$: Lower Limit confidence interval. CI$_{UL}$: Upper Limit confidence interval. SD: Standard Deviation. Min. and Max.: Minimum and Maximum. Q1 and Q3: quartiles 1 and 3. N$_{>5\%}$ = number of times the index exceeded 5%. N$_{>10\%}$ = number of times the index exceeded 10%.

These results can be seen graphically in Figure 3, where the results have been classified according to the result obtained in the DI. On the x-axis we find the percentage values, while inside the bars we can observe the absolute values. The highest rate of successful reproduction (DI < 5%) was found when the alpha coefficient was reproduced (97.55%), followed by the lower limit of the confidence interval of the alpha coefficient (95.79%). However, the lowest reproducibility rate was found in the heterogeneity indices, specifically in the Q statistic (64.79%).

Figure 3.

Reproducibility results for the discrepancy index.



Note. The x-axis shows the percentage of the total number of reproductions for each variable. The internal values show the absolute values of each category of DI. LL = Lower Limit of the confidence interval. UL = Upper Limit of the confidence interval.

The summary of the results of the correlations between the reported and reproduced values are shown in Table 3. It is evident that, in all cases, the values of all correlations are above .8 and statistically significant.

Table 3.

Correlation results between reported and reproduced values.

		95% CI					
	r_{xy}	**LL**	**UL**	R^2	*t*	*df*	*p*
Alpha	.868	.824	.901	.753	22.151	161	<.001
CI_{LL}	.901	.854	.933	.811	20.001	93	<.001
CI_{UL}	.889	.838	.925	.791	18.736	93	<.001
I^2	.823	.728	.887	.678	11.777	66	<.001
Q	.988	.981	.992	.976	52.792	69	<.001

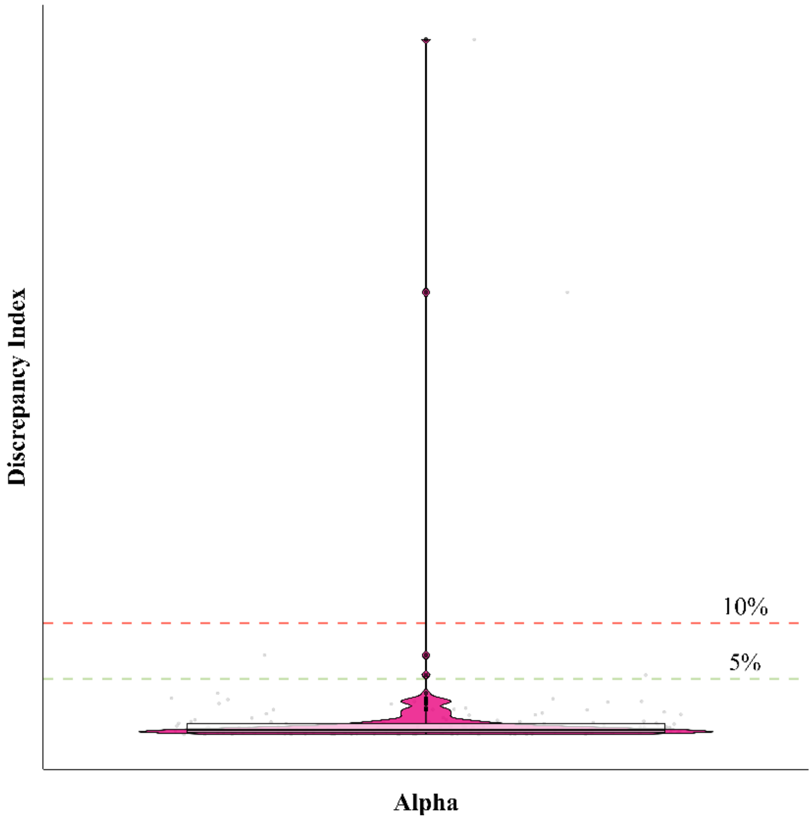
Note. CI_{LL}: Lower Limit confidence interval. CI_{UL}: Upper Limit confidence interval. LL = Lower Limit confidence interval of r_{xy}. UL = Upper Limit confidence interval of r_{xy}

Alpha Coefficient

DI results (Table 2) obtained by reproducing the average alpha coefficients showed that only in four occasions the discrepancy rate between the reported and reproduced values exceeded the 5% limit (in two occasions, 10%). Although the maximum discrepancy value was 62.313, in 159 occasions (97.55%) the *DI* was below 5%, which means successful reproduction. Figure 4 shows graphically the distribution of the Discrepancy Indices.

Figure 4.

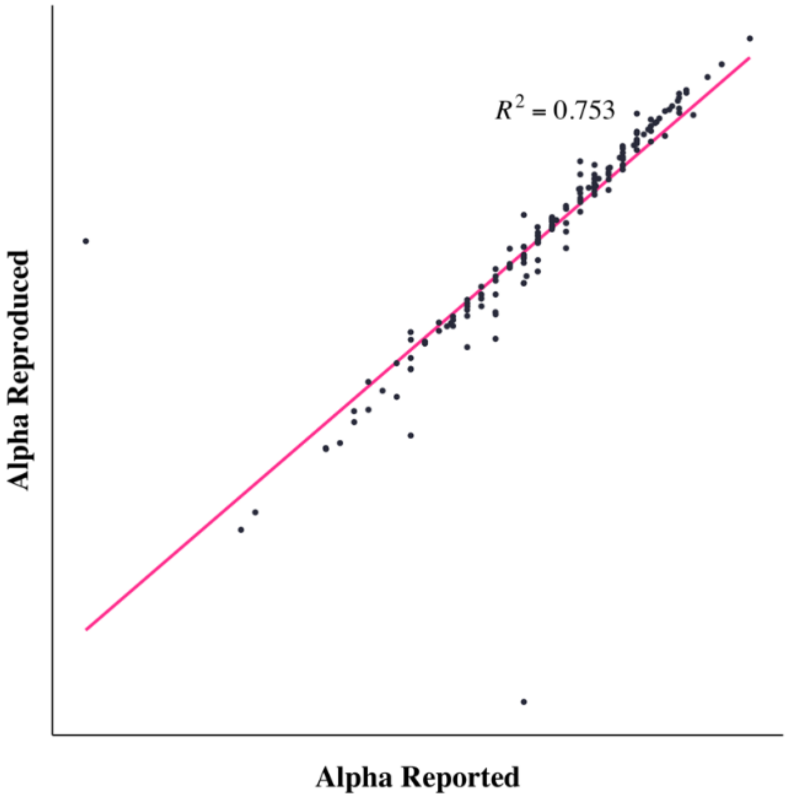
Violin plot of the distribution of the discrepancy index for the mean coefficients. Above 5%, the differences were significant and beyond 10%, replication had been unsuccessful.



Moreover, concerning the results of the correlation between the reported values and the reproduced values (Table 3), we found a statistically significant correlation (.868). Figure 5 shows graphically the situation between reported and reproduced values, as well as the value of the coefficient of determination (R^2 = .753).

Figure 5.

Scatter plots of reported and reproduced alpha coefficients. The coefficient of determination of the relationship between the two coefficients has also been calculated.

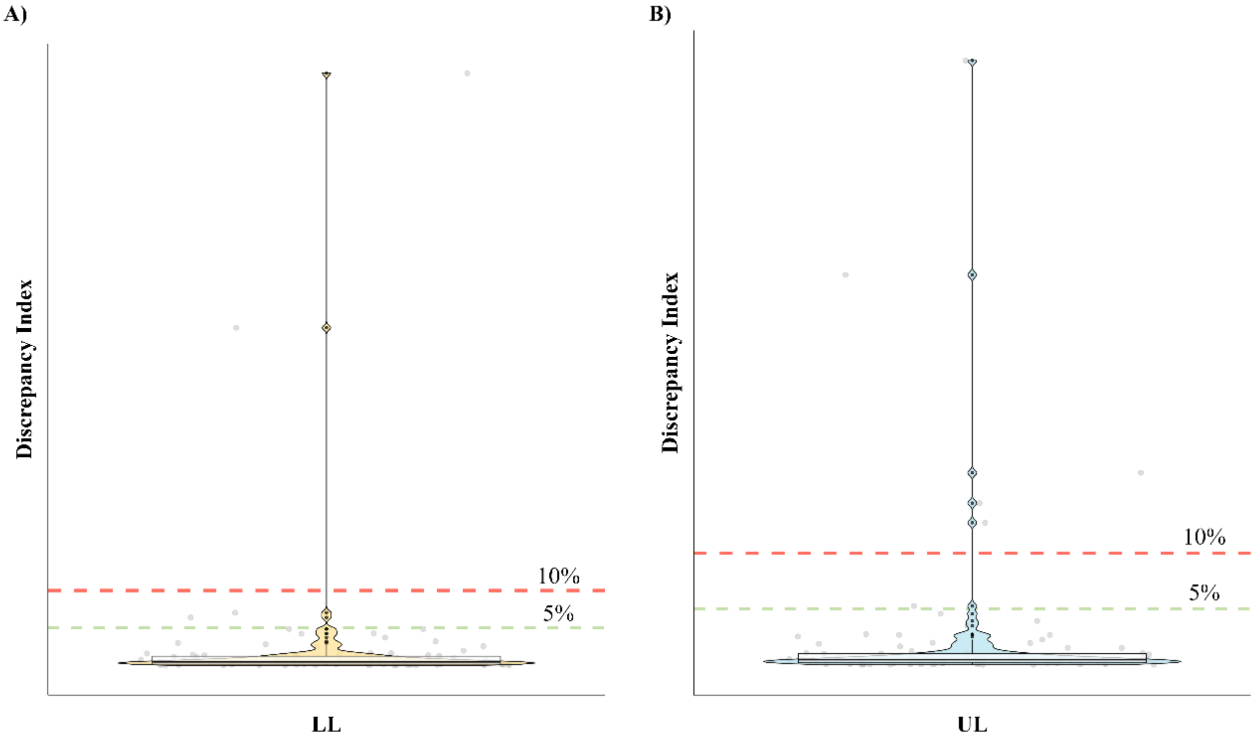


Confidence Intervals

Ninety-five confidence intervals of the average coefficient were recalculated. Results of the DI between the reported and reproduced values are shown in Table 2. For the lower limit, we found that the 5% limit was exceeded in four occasions, two of them being greater than 10%. On the other hand, with respect to the upper limit, the value of 5% was exceeded in six occasions, five of them being greater than 10%. Figure 6 shows graphically the distribution of this index, both for the lower limit (A) and the upper limit (B).

Figure 6.

Violin plot of the distribution of the discrepancy index for both limits of confidence interval of the average coefficients. From 5% onwards, the differences were noticeable and beyond 10%, the reproduction had been unsuccessful.

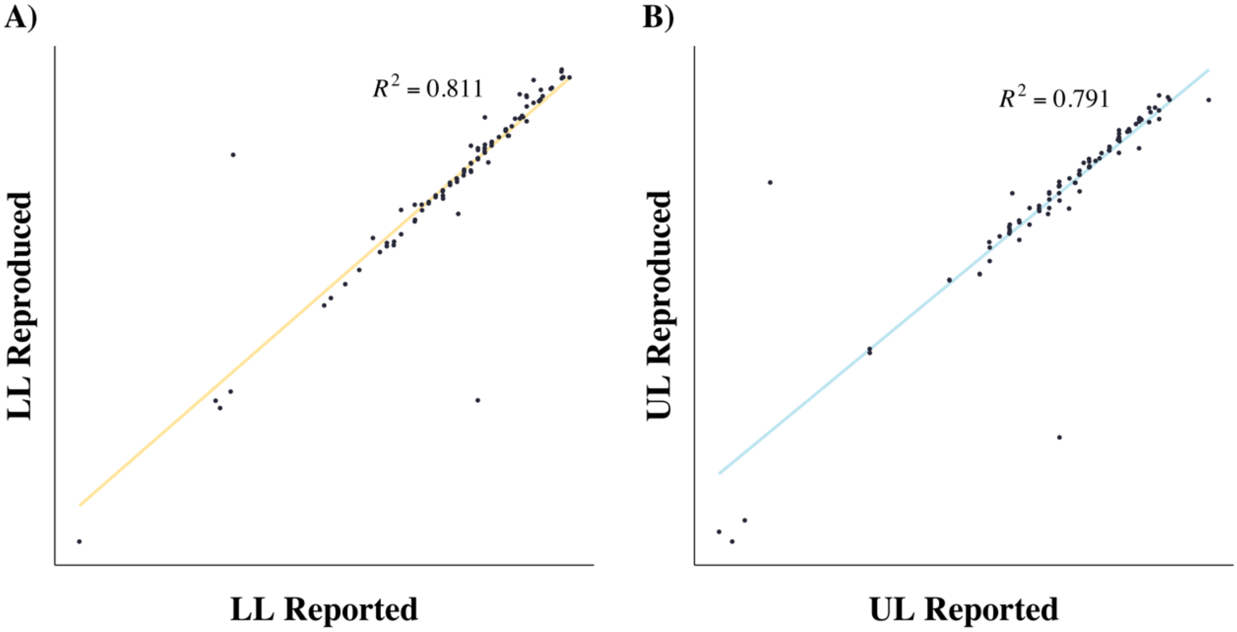


Therefore, referring to the classification established to determine reproduction success, in 91 occasions (95.79%) of the lower limit and 89 (93.68%) of the upper limit, reproduction was practically perfect. Two occasions, for the lower limit, were considered as reproduction error (2.11%), and five (5.26%) for the upper limit.

The correlation results can be found in Table 3. Statistically significant results have been observed for both limits (r_{xy_LL} =.901; r_{xy_UL} =.889). Figure 7 shows this relationship graphically for both the lower (A) and upper (B) limits, as well as their coefficients of determination ($R^2_{LL} = .811$; $R^2_{UL} = .791$).

Figure 7.

Scatter plots of reported and reproduced lower (A) and upper (B) limit of the confidence interval of the average coefficient. The determination coefficient of the relationship between the reported and reproduced lower limit of the confidence interval has also been calculated.



Heterogeneity Indices

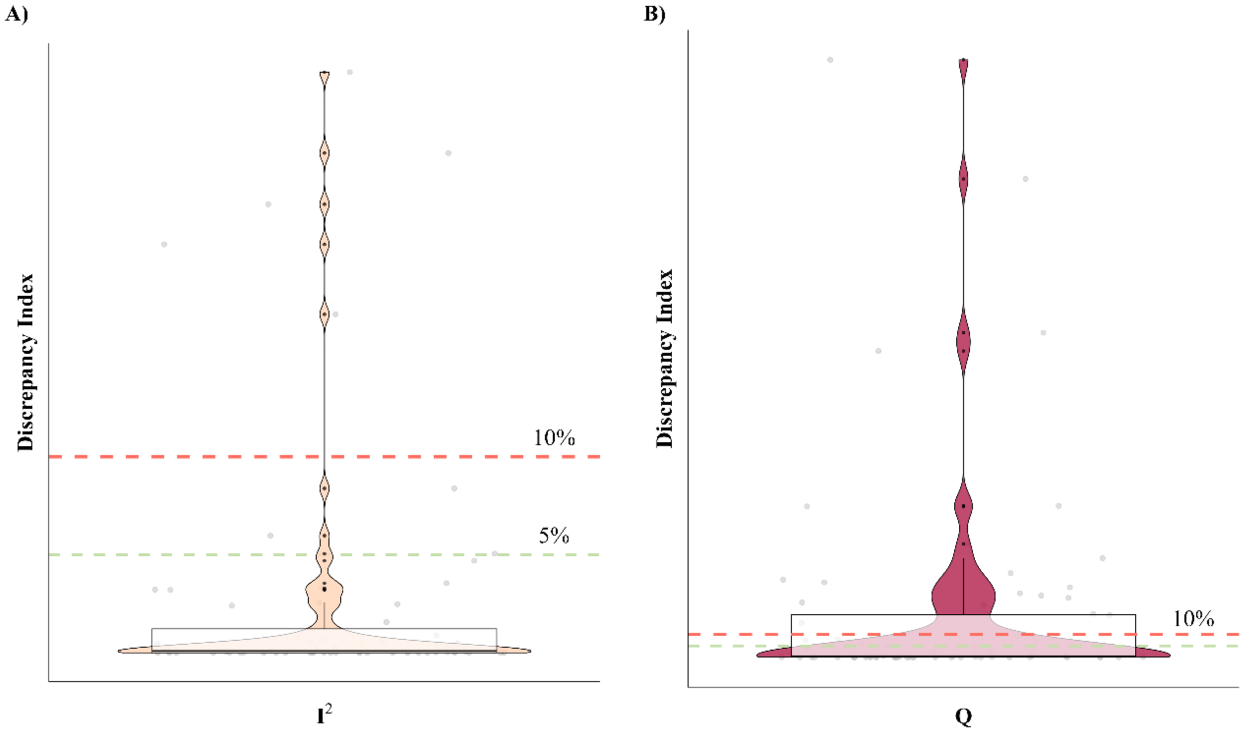
The most notable case of underreporting has been in the case of heterogeneity indices: in more than half of the meta-analyses none of the indices were reported.

The discrepancy index limit was also exceeded most often in this variable: in eight occasions the I^2 index showed values above 5% (five of them above 10%), and as regards the Q statistic, the limit was exceeded in 25 occasions (18 above 10%). Employing the DI classification, in 60 occasions (88.24%) a successful reproduction of the I^2 index was achieved and in 46 occasions (64.79%) of the Q statistic. It should be noted that in the Q statistic the percentage of non-reproduced results was over 25% (N_Q = 18; 25.35%). Figure 8 shows graphically the distribution of these results.

.

Figure 8.

Violin plot of the distribution of the discrepancy index for (A) I^2 index and (B) Q statistic. From 5% onwards, the differences were noticeable and beyond 10%, the reproduction had been unsuccessful.

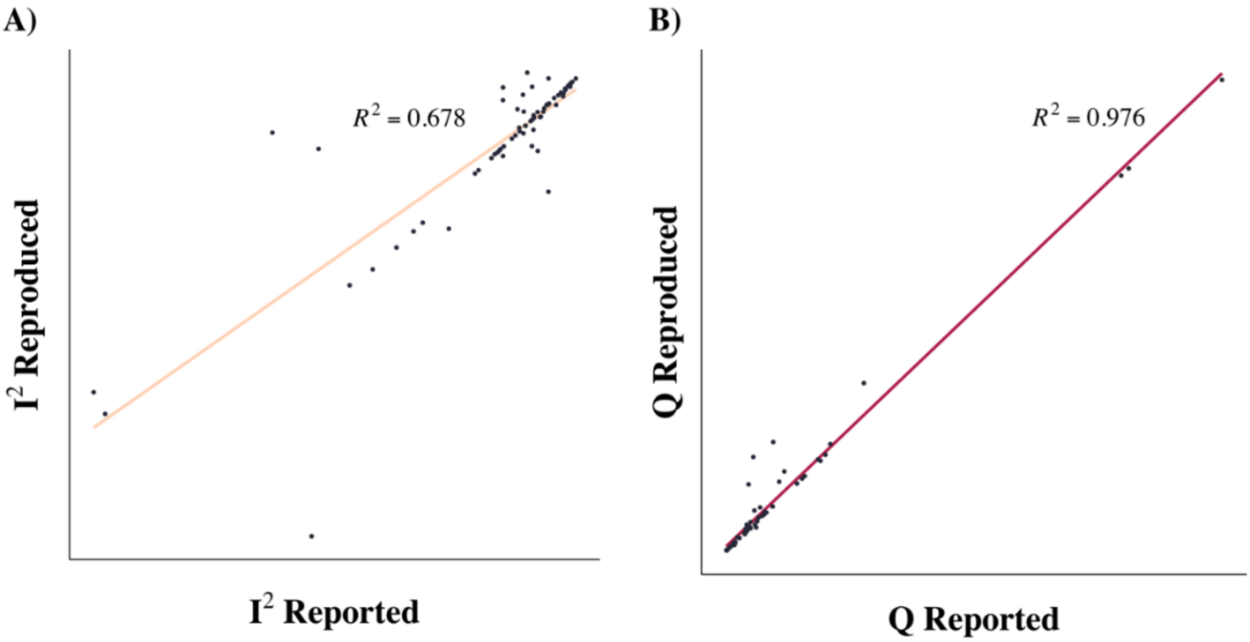


As we observed in Table 3, both indices presented statistically significant correlations with quite strong results, mainly in the case of the *Q* statistic, which showed a practically perfect correlation (r_{xy_I} = .823; r_{xy_Q} = .988). Figure 9 also shows the coefficients of determination for both indices ($R^2_{I^2} = .678$; $R^2_Q = .976$).

.

Figure 9.

Scatter plots of reported and reproduced I^2 index (A) and Q statistic (B). The coefficient of determination of the relationship between the reported and reproduced upper limit of the confidence interval was also reported.



3.4 Discussion

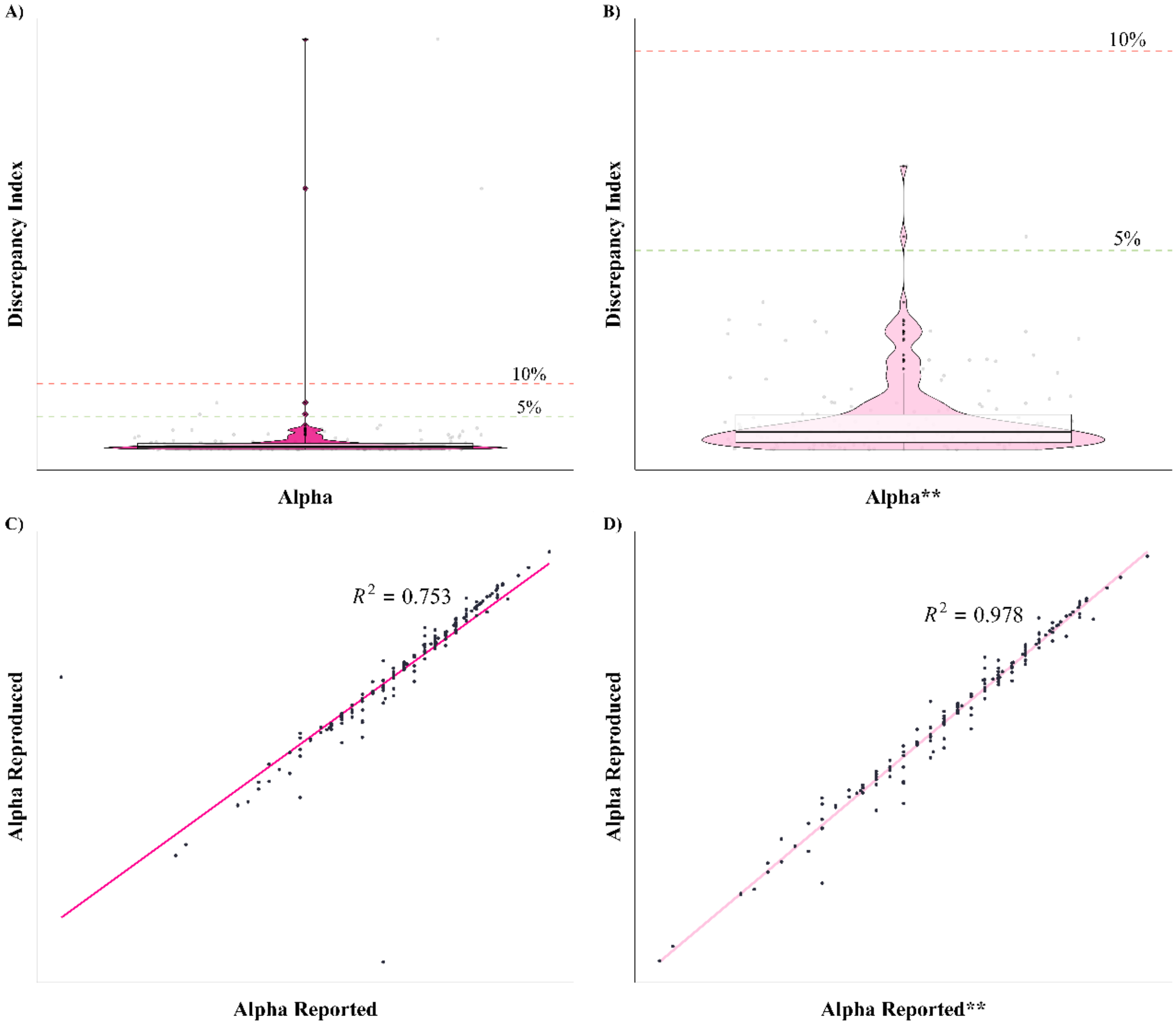
In this research we have conducted a reproducibility study of reliability generalization meta-analyses. In particular, we reproduced 163 average alpha coefficients and 95 confidence intervals (CI) of that coefficient. Moreover, we recalculated the heterogeneity indices: 68 I^2 indices and 71 Cochran's *Q* statistics. To evaluate the results obtained with respect to the original reported in the literature, we established two analytic strategies: first, we calculated a discrepancy index (*DI*) that indicated the percentage in absolute value of difference between a reported result and the reproduced one (Equation 1); and secondly, we calculated the correlation between the reported and the reproduced values. In order to establish whether the difference was significant or not, we classified the *DI* result into three categories: if the result was below 5%, the reproduction was considered successful; secondly, if the value exceeded 5% but not 10%, we considered the reproduction result to be approximate; and, finally, if the value exceeded 10%, the reproduction failed and there were notable differences between the reproduced and reported values. In terms of the results of the correlations, to consider that the reproduction was successful, these were expected to be above .80 and statistically significant.

As seen in the previous section, our results showed that for all the variables studied (α, CI, I^2 and Q) the correlation between the reported and reproduced values was higher than .80 (Table 3). However, we found much more variability when interpreting the DI values (Table 2). For the average coefficient and its confidence interval, the *DI* was below 5% almost 95% of the cases (α = 97.55%, CI$_{LL}$= 95.79%, CI$_{UL}$= 93.68%), whereas, for the heterogeneity indices, this rate was reduced (I^2 = 88.24%, Q = 64.79%).

Overall, these results showed a positive outlook, since, with the exception of the Q statistic, the fact that the variables studied exceed 80% of successful reproductions is not usual in this type of study (Artner et al., 2021; Hardwicke et al., 2018; Maassen et al., 2020). Attempting to explain why the total number of reproductions has not been reached, we have detected an error produced on a specific scale. In this work, two subscales showed the results of the meta-analysis for each subscale changed between them. The problem was found to come from the database itself, where each primary study was classified according to the version of the test used (some primary studies had used different test versions than those specified in the database). Without identifying the specific studies that this error affected, removing the results of these two subscales that clearly showed erroneous data, we found that, for example, the correlation between the average reported and reproduced coefficients changed from $r_{xy}^2 = .868$ to $r_{xy}^2 = .989$ (Figure 9). The discrepancy rate was also reduced for average alpha, with only two studies having between 5 and 10% discrepancy. This change in the results was common to all the variables considered.

Figure 9.

Violin and scatter plots of reported and reproduced average alpha coefficients. The violin plots show the distribution of the discrepancy index for (A) all the average alpha coefficients and (B) average alpha coefficients without the erroneous subscales. The scatter plots show the correlation between average alpha (C) with all the subscales and (D) without the erroneous subscales. The coefficient of determination of the relationship has also been calculated.



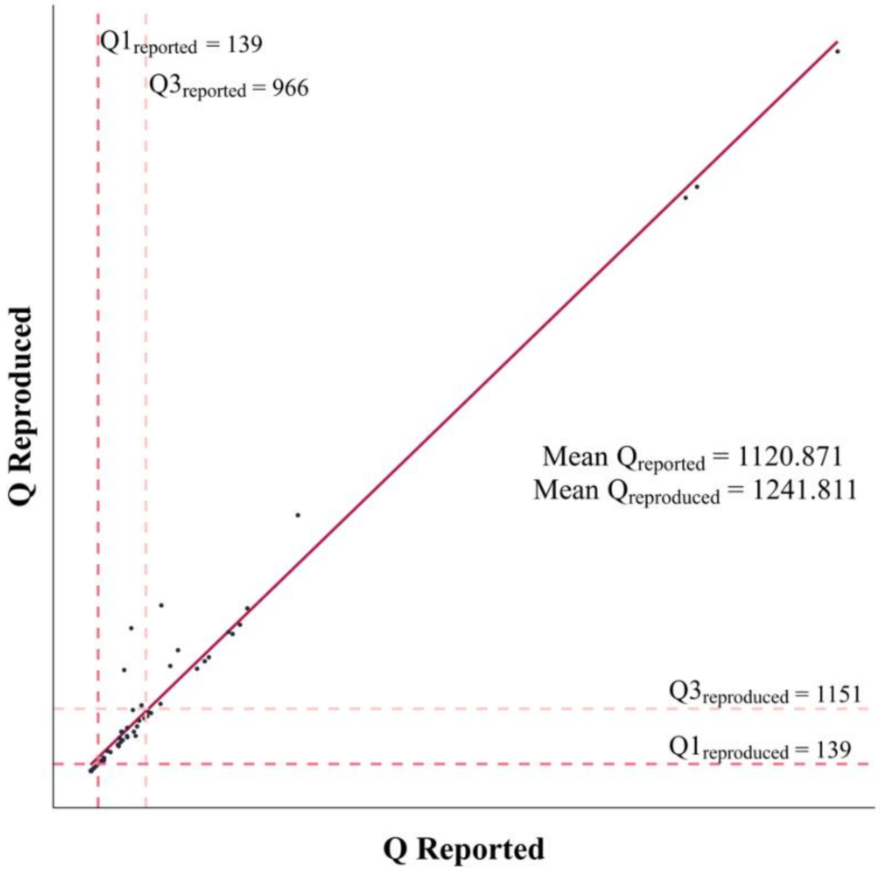
Note. ** = Average alpha coefficients without the two erroneous subscales.

Despite this error being an anecdotal case, which showed its influence on the results, the lack of reproducibility that we have found, especially in the heterogeneity indices, cannot be explained solely by this error. Focusing on the case of the *Q* statistic, one plausible explanation for this phenomenon is that the values that *Q* takes are between 0 and +∞, thus implying that the results of *Q* tend to be much higher than other variables that we have tested such as average alpha, whose values are between 0 and 1. Furthermore, we have to be aware that the data have context and, in this case, the value of the *Q* statistic provides information about the presence of heterogeneity, together with its associated probability value. In a hypothetical case where $Q_{reported}$ = 790 and $Q_{reproduced}$ = 700 the result of its DI would exceed 10%. However, from a substantive point of view, this difference does not affect the conclusion to be drawn from it. In other words, the result remains congruent. Something similar happens with the I^2 index, although in this case the results of the DI are more positive ($DI_{<5\%}$ = 88.24%). We might think that this discrepancy index works best when the values it evaluates are small and clearly bounded. With another strategy for analysing the results, we can see that this explanation makes sense, since the Pearson correlation for the reported and reproduced values of the *Q* statistic has been r_{xy} =.988 and for the I^2 index, r_{xy} = .823, both statistically significant.

Figure 10 shows a scatter plot of the reported and reproduced *Q*-values including quartiles 1 and 3 of both groups of values, as well as their averages. From this figure we can see that the distributions of the reported and reproduced values are quite similar, with the values corresponding to the reproduced group being slightly higher at quartile 3 and at the mean. Though not included in the figure, the median values were also very similar ($Q_{reported}$ = 496.14, $Q_{reproduced}$ = 486.58). With all this we can deduce that using more than one interpretation strategy makes both complement each other, giving a more complete view of the phenomenon under study.

Figure 10.

Scatter plots of reported and reproduced Q index. This plot shows the correlation, the quartiles 1 and 3 and the mean value for each group of values.



Note. Q1 = quartile 1. Q3 = quartile 3.

It is important to note that, while the results showed a good replication rate for each variable, the reporting rate could be improved. For example, of the 152 studies initially collected, 92 studies (60.53%) did not report the database with the primary studies, so had to be directly discarded. From the 170 average coefficients reported, only 60% of them also provided the value of the confidence interval, 42% the value of the I^2 index and 44% the value of the *Q* statistic. In some cases, these indices may not have been reported because the researchers assumed a fixed-effect model. Nevertheless, the fixed-effect model has not been the most common (Figure 1) and it is still recommended to verify that the assumed model is correct by assessing the heterogeneity of the component studies.

The issue of information reporting has also been recently studied in the study of Sánchez-Meca et al. (2021). Specifically, they elaborated a systematic review of 150 RG meta-analysis studies applying the REGEMA guideline (Sánchez-Meca et al., 2021). Of the 30 items that compose this guide, in 12 of them the reporting rate was less than 50%. The use of such guidelines is helpful in facilitating the reporting and improving the transparency of experimental studies. These guidelines specify what information must be reported in order for the study to be fully transparent and reproducible or replicable. For systematic reviews or meta-analyses, the most widely used and most popular guideline is the PRISMA guideline (*Preferred Reporting Items for Systematic reviews and Meta-Analyses*) (Liberati et al., 2009), while, in the case of reliability generalization meta-analyses, the REGEMA (*Reliability Generalization Meta-Analysis)* checklist (Sánchez-Meca et al., 2021) is particularly designed for this kind of meta-analysis. Additionally, pre-registration tools, such as PROSPERO, also help to maximise information reporting and improve transparency practices.

This lack of good practice in terms of transparency and reporting means that research results are not completely reliable and that these results are not a good representation of real phenomena. Publicising materials, tools, datasets, scripts and, in fact, everything necessary to reproduce the results obtained in a scientific study, should be imperative.

It is also important that all available materials are easily understandable, as it is often not only a lack of information but also a lack of clarity that hinders the reproduction process. Indeed, some of the errors identified throughout the research have been related to confusing explanations of the process of analysing and obtaining the results, and to the appearance of different values in different parts of the same article (for example, the sample size of the primary studies changed from one table to another, or the meta-analytical results were different in the text and in the table). Some recommendations for improving transparency and open science practices can be found in (Artner et al., 2021).

3.4.1 Limitations and Future Research

Despite the fact that the results we have found are overall positive, it should not be forgotten that we were able of analysing 32 meta-analytic studies out of the 152 that met our initial criteria. In other words, we only recalculated the meta-analytic results of 20% of the studies that met our inclusion criteria in the first place. A limitation of this work may have been the lack of contact with the corresponding authors to request such primary data.

Updating the search and completing the range to 2023 is the next objective, with a deeper revision of the errors in reporting and lack of transparency of meta-analyses, as well as studying whether there is a real relationship between improved research practices and the year of publication of the studies. Broadening the type of coefficients allowed in the inclusion criteria is also an objective of the upcoming research.

3.5 Conclusion

In this study, we have evidenced that the reliability generalization meta-analyses published to date show replication rates between 98 and 65%. Moreover, the correlation rate between reported and recalculated results is above .80 for all parameters evaluated. This result is positive with respect to the results of previous studies on other types of meta-analysis, giving solidity and credibility to this type of work. However, there is still a long way to go in terms of transparency and open science in this area. It seems important to point out that both the researchers themselves and the publishers of the scientific journals where these studies are published are aware of the importance of providing the necessary materials to corroborate the results obtained and the conclusions drawn by them. Without a good report of the method established to carry out the analyses of the work and without publicizing in some way the materials used, it is impossible to carry out this type of work that can corroborate that the meta-analysis of reliability generalization is a useful tool that provides robust results.

Chapter 4

Study 3:

“The Reliability Generalization Meta-Analysis through the multilevel approach: A comparison between the traditional technique and the multilevel perspective”

4.1 Introduction

In the previous chapters it was seen that the meta-analysis is considered the best tool for synthesising and integrating empirical results. Despite all its advantages, the application of this tool also has some drawbacks. Focusing exclusively on the meta-analysis of reliability generalization (RG), on one side, by relying on primary studies, it is not possible to carry out an RG study with the most suitable coefficients for each instrument, but only to carry out the meta-analysis with the coefficients provided by these studies. For example, although the use of Cronbach's alpha coefficient to determine the reliability of questionnaires, especially those with a multidimensional nature, has recently been questioned, if the primary studies do not provide the empirical data from their research, or only apply this coefficient, without consider more robust coefficients (such as omega coefficients), the meta-analysis cannot obtain as accurate results as one would wish. On the other hand, another drawback of the conventional procedure of conducting a meta-analysis is the way in which it deals with the dependency relationships that may arise. For example, a test that has different subscales and these have been applied to the same group of participants will show some dependence between them, especially considering that what these subscales assess are different dimensions within the same psychological construct. It is common in sciences such as Psychology for data to have a hierarchical structure, thus introducing dependency into the data. Hierarchy must be considered when performing the relevant statistical analyses that assume the independence of the residuals (Fernández-Castilla et al., 2019, 2020).

As mentioned in the introduction to this thesis, when a test has a multidimensional structure with different subscales, all forming part of the same psychological construct and, in addition, these scales are applied to different groups within the same scientific study, dependence between scores can arise. This is especially important as the observed effect size may be providing different information than expected by the model (Assink & Wibbelink, 2016; Van den Noortgate et al., 2013). When a reliability generalization meta-analysis is performed in the conventional method, this dependence is not considered, because if a test has a multidimensional structure, each of the dimensions is analysed in separate meta-analyses, thus abolishing any dependence that may exist between scores. This practice is not particularly suitable because by selecting only a part of the available data, the results obtained are less accurate and the statistical analyses are less powerful (Assink & Wibbelink, 2016; Van den Noortgate et al., 2013, 2015). Moreover, in many cases, by taking only a part of the sample for each subscale, it is impossible to meta-analyse these coefficients, as they have an insufficiently small number of observations.

The best way to deal with the problem of dependency is to try to model it. Raudenbush et al. (1988) proposed the multivariate model for analysing multivariate effect sizes. This model uses the estimated covariance matrix of the multivariate effect sizes, which makes it possible to use all available information in a single analysis and estimate the treatment effect for each dependent variable. However, the application of this model to the actual data found in primary studies is complicated, since to obtain such a matrix it is also necessary to have the correlations between the dependent variables, which are rarely reported in primary studies, as is the case with the raw data, which are also often not available to the meta-analyst. Fortunately, the rise of open science is changing things.

An alternative way to model dependence in meta-analyses is the application of multilevel models. Raudenbush and Bryk (1985) were the first to propose the use of multilevel models to perform meta-analyses. The major advantage of this model is that it uses all relevant effect sizes, thus preserving all the information provided by the primary studies (Assink & Wibbelink, 2016). That is, it is not necessary that all studies report exactly the same results. Moreover, this model is very flexible, as the data can be structured in the most convenient way and the model will adapt to it. Another advantage is that it automatically accounts for the hierarchical structure of the data when entering the analysis (Van den Noortgate et al., 2013).

It seems that the three-level structure of the multilevel model is the best method to deal with dependence (Assink & Wibbelink, 2016; Van den Noortgate et al., 2013, 2015). This structure states that the variance components are distributed over three levels: the first level refers to the sample variance of effect sizes (or reliability coefficients), the second level is the variance between effect sizes drawn from the same study, and the third level is the variance between studies.

Considering the type of data, we are dealing within a meta-analytic reliability generalization study, there is nothing to indicate that there are no dependence relationships between the scales if they have been administered to the same sample of participants. Hence, it is of particular interest to apply such an approach to an RG study. Therefore, the main objective of this study was to test whether there are differences between the application of a meta-analytic model of reliability generalization from the traditional approach and from the multilevel approach. To our knowledge, whenever multilevel models have been applied in a study of this nature, the subscales have been analysed independently, not as part of the same analysis (e.g. Maes et al., 2015).

Another model also employed to model dependency by focusing on standard error adjustment is the Robust Variance Estimation (RVE) model, proposed by Hedges et al. (2010). One of the advantages of this model is that it does not require information on the covariance structure of the effect size estimates. The RVE model establishes two types of dependence in meta-analyses: the first one called "correlated effect sizes" refers to multiple effect size estimates reported by a primary study, when the different underlying measures are correlated or even when the same control group has been used for all treatment groups. The second type of dependency has been called "hierarchical effect size" and occurs when the same researchers publish multiple studies on the same topic but using different population samples, including when the reports are due to independent experiments (Pustejovsky & Tipton, 2022; Tanner-Smith & Tipton, 2014). Another advantage of the RVE model is that it is not particularly relevant to correctly identify in the analysis whether it corresponds to the hierarchical or the correlational method, as the results do not vary substantially between one or the other method (Tanner-Smith & Tipton, 2014).

The main difference between multilevel meta-analysis and RVE is the way heterogeneity is addressed. Unlike the multilevel model, the RVE model uses the heterogeneity parameters to estimate the inverse variance weights. Heterogeneity is only incidental to RVE (Tanner-Smith & Tipton, 2014).

4.1.1. Current Study

The main objective of this work was to study how different meta-analytic perspectives behave with the same set of scales. On the one hand, we were interested in testing whether there were differences between the application of multilevel models in different conditions and the application of the conventional procedure, which separates each subscale into independent meta-analyses. These different conditions were stipulated in terms of the number of subscales and the number of primary coefficients per subscale. Two other aspects we compared were the application of Bonett's transformation (2002) to the coefficients and the improved Knapp-Hartung's method (2001) for the calculation of confidence intervals. Including a transformation of the coefficients and an alternative method for the construction of the confidence intervals introduces a minor sensitivity analysis to the results obtained procedurally. To facilitate the interpretation of the results, a discrepancy index has been calculated to check the percentage change we found between them.

Table 1 presents the different comparisons carried out to calculate the discrepancy index: 4 comparisons have been established for the calculation of the average coefficient and 5 comparisons for the calculation of the confidence width.

Table 1.

Study comparisons for average coefficient and confidence width

	Average alpha coefficient	Confidence Width
Conventional vs Multilevel	✓	✓
Conventional vs Homo/Heteroscedastic	✓	✓
Homoscedastic vs Heteroscedastic	✓	✓
Untransformed vs Transformed	✓	✓
Standard vs Knapp-Hartung's	-	✓

4.2 Method

4.2.1. Study Selection Criteria

To test all these analytical approaches, four different scales were selected that had been meta-analysed by our research group. This was decided in order to have as much information as possible. These four scales were selected according to two criteria: the number of subscales that made up the scale itself (many vs. few) and the number of observations in each of the subscales (many vs. few). The aim of this selection was to observe how the different methods performed in different situations, trying to identify if there was any pattern in the results. With these two parameters, we selected the 4 scales shown in Table 2.

Table 2.

Selected scales

		Number of subscales	
		1-3	**> 4**
Number of observations (primary coefficients)	**< 20**	FOCI	PI-R
	> 21	CAPS	DOCS

4.2.2. Scales Included

FOCI (Florida Obsessive-Compulsive Inventory)

The Florida Obsessions and Compulsions Inventory (FOCI; (Storch et al., 2007) allows us to examine both obsessive-compulsive symptoms and their severity in a self-report format. It is composed of two subscales: the *Symptom Checklist*, which assesses the presence or absence of ten common compulsions in this disorder; and the *Symptom Severity* consisting of 5 items in Likert-type scale with 5 options (from 0 to 4). Table 3 shows a summary of the reliability results (Sandoval-Lentisco et al., 2023).

Table 3.

Summary of FOCI's psychometric properties obtained in Sandoval-Lentisco et al. (2023)

			95% CI		**95% Cr. I**			
Subscales	***k***	***α***	**LL**	**UL**	**LL**	**UL**	***Q***	***I*²**
FOCI Checklist	17	.826	.815	.838	.794	.859	26.243	40.57
FOCI Severity	15	.882	.861	.903	.816	.948	118.633**	90.04

Note: k = number of studies; α = mean coefficient alpha; *LL* and *UL* = lower and upper limits of the 95% confidence and credibility interval for α; Q = Cochran's heterogeneity Q statistic; I^2 = heterogeneity index. **p<.0001.

PI-R (Padua Inventory-Revised)

The Padua Inventory Revised (PI-R; Van Oppen et al., 1995) is a scale that assesses obsessive-compulsive symptomatology and is composed of 41 items grouped into 5 subscales: impulses (7 items), washing (10 items), checking (7 items), rumination (11 items) and precision (6 items). Each item is answered in Likert-type scale with 5 options (0-4). Table 4 shows a summary of the reliability results (Núñez-Núñez et al., 2022).

Table 4.

Summary of PI-R' psychometric properties obtained in Núñez-Núñez et al. (2022)

			95% CI		95% PI			
Subscales	***k***	***α***	**LL**	**UL**	**LL**	**UL**	***Q***	***I*²**
Whole Scale	28	.92	.91	.93	.85	.96	458.306**	94.6
Impulses	19	.79	.76	.82	.65	.87	176.122**	90.6
Washing	20	.89	.86	.91	.68	.96	777.177**	97.6
Checking	18	.88	.86	.89	.80	.93	175.537**	91.4
Rumination	19	.87	.85	.89	.75	.93	313.529**	94.6
Precision	18	.74	.69	.77	.49	.86	224.948**	94

Note: k = number of studies; *α* = mean coefficient alpha; *LL* and *UL* = lower and upper limits of the 95% confidence/prediction intervals for *α*; *Q* = Cochran's heterogeneity *Q* statistic; I^2 = heterogeneity index. ***p*<.0001

CAPS (Child and Adolescent Perfectionism Scale)

The Perfectionism Scale for Children and Adolescents (CAPS; Flett et al., 2016) is the most widely used instrument to assess perfectionism in children over 8 years of age. It consists of 22 items on a Likert-type scale with 5 response options and two subscales: Self-Oriented Perfectionism (12 items), which assesses the motivation and effort to be perfectionist and the tendency to self-criticism; and Socially Prescribed Perfectionism (10 items), which assesses beliefs about the demands of perfectionism from the environment. In Table 5 shows the reliability results of the scale (Vicent et al., 2019).

Table 5.

Summary of CAPS' psychometric properties obtained in Vicent et al. (2019)

			95% CI		95% P I			
Subscales	***k***	***α***	**LL**	**UL**	**LL**	**UL**	***Q***	***I*²**
Whole Scale	11	.87	.84	.90	.73	.94	174.97**	96.8
SPP	51	.84	.82	.85	.72	.91	851.738**	93.4
SOP	47	.83	.81	.84	.66	.91	1010.134**	95

Note: k = number of studies; *α* = mean coefficient alpha; *LL* and *UL* = lower and upper limits of the 95% confidence and prediction intervals for *α*; *Q* = Cochran's heterogeneity *Q* statistic; I^2 = heterogeneity index. ***p*<.0001.

DOCS (Dimensional Obsessive-Compulsive Scale)

Dimensional Obsessive-Compulsive Scale (DOCS; Abramowitz et al., 2010) is composed of 20 items structured in 4 dimensions -5 items each-: *contamination, responsibility for harm, unacceptable thoughts, and symmetry*. This scale assesses both the presence of symptoms and the distress caused by such symptomatology. Items are evaluated on a Likert-type scale with five options, where scores on each subscale can range from 0 to 20, and from 0 to 80 on the full scale. Table 6 shows a summary of the reliability properties (López-Nicolás et al., 2021).

Table 6.

Summary of DOCS' psychometric properties obtained in López-Nicolás et al. (2021)

			95% CI			
Subscales	**k**	**α**	**LL**	**UL**	**Q**	**I²**
Whole Scale	72	.925	.92	.931	567.646**	91.34
Contamination	58	.881	.863	.899	2475.88**	98.32
Responsibility	50	.905	.893	.917	996.208**	96
Unacceptable thoughts	51	.913	.904	.922	641.066**	93.97
Symmetry	49	.914	.906	.922	481.511**	91.83

Note: k = number of studies; *α* = mean coefficient alpha; *LL* and *UL* = lower and upper limits of the 95% confidence interval for *α*; *Q* = Cochran's heterogeneity *Q* statistic; I^2 = heterogeneity index. ***p*<.0001

4.2.3. Data analysis

To carry out the conventional meta-analysis, separate meta-analyses were conducted for each subscale, as indicated by the authors in the original RG meta-analyses. All analyses were computed in R with the *metafor* package (v3.4-0; Viechtbauer, 2010). Parameters were estimated using the restricted maximum likelihood estimator (REML). The multilevel models were arranged with a 3-level structure: the sample variance of the coefficients at level 1, the variance between coefficients within each study at level 2 and the variance between studies at level 3. The intercept was not included in the model because the interest was not in the overall effect. The structuring of the data and the computational script were developed following the tutorial of Assink and Wibbelink (2016). All multilevel model analyses were also calculated with the same software, package and estimator as the conventional analyses. On the other hand, to calculate the RVE model, the weighting method used was the *correlated effects*, and the analyses were carried out with the R package *robumeta* (v2.0; Fisher & Tipton, 2015).

In order to establish comparisons and to be able to determine whether these are significant or not, we calculated a discrepancy index by setting one of the values as a reference value and comparing the rest of the results with it. This reference value has been changing according to the objective comparison. The formula applied to calculate this index was:

$$D(Alpha)_j = \left(\frac{\alpha_j - \alpha_c}{\alpha_c}\right) \times 100,$$

Where α_j refers to the value we want to compare with respect to α_c which is the value we define as reference.

To establish the differences between the results in the calculation of the average coefficient, 4 different discrepancy indices were calculated.

The first comparison was between the conventional model and the multilevel model (both raw and transformed coefficients). In this case, the reference value was the average coefficient obtained using the conventional model. The second comparison was between the conventional model and the homo- and heteroscedastic multilevel model (both raw and transformed coefficients). The reference value was the average coefficient obtained using the conventional model. The following comparison was performed between the two multilevel models: homo- and heteroscedastic model (both raw and transformed coefficients). The reference value used was the average coefficient obtained using the homoscedastic model. And finally, the last comparison was between the average coefficient with raw coefficients and the average coefficient using Bonett's transformation (within each model). The last reference value was the untransformed average coefficient in each model.

To establish the differences between the results in the calculation of confidence intervals, 5 different discrepancy indices were calculated.

The first discrepancy index was calculated by comparing the conventional model with the multilevel model (both using the improved Knapp-Hartung's formula for calculating confidence intervals or the standard method and using raw or transformed coefficients). This index used the confidence width obtained through the conventional model as the reference value. The second index was calculated by comparing the conventional model and the homo- and heteroscedastic multilevel model (both using raw and transformed coefficients and applying the standard or Knapp-Hartung's method). The reference value used was the confidence width obtained using the conventional model. The next index was estimated by comparing homo- and heteroscedastic multilevel model (both using raw and transformed coefficients and applying either the standard or the Knapp-Hartung's method). This index set as the reference value the confidence width obtained using the homoscedastic multilevel model. The fourth index was generated by comparing the results of the confidence width calculation using Knapp-Hartung's or the standard method (within each model and including the comparison between raw and transformed coefficients). In this case, the reference value was the confidence width obtained using the standard method. Finally, the last discrepancy index was calculated by comparing the impact of using Bonett's transformation on the average coefficient to calculate confidence intervals (within each model and including the comparison with Knapp-Hartung's versus the standard method). This index used as reference value the confidence width obtained with the untransformed average coefficient in each model.

Tables 4A.1-4A.6 in Appendix 4A summarise the different combinations that were compared both for the calculation of the alpha coefficient and for the calculation of its confidence interval. The complete dataset as well as the script codes used to analyse them are openly available at: http://bit.ly/40eykLj

4.3 Results

4.3.1 Characteristics of the Scales

The 17 subscales were extracted from 4 complete scales. These subscales had a number of primary coefficients (*k*) that ranged between 14 and 72. Table 7 contains the results of median, mean, SD, minimum and maximum, and first and third quartiles of alpha coefficient and sample size for each subscale.

Table 7.

Descriptive data for each scale and subscale

Scale	Subscale		Mean	SD	Min	Q1	Median	Q3	Max
DOCS	Whole Scale	α	.92	.029	.8	.91	.93	.94	.97
		N	281.82	278.32	16	87	201.5	391.5	1299
	Contamination	α	.875	.075	.61	.83	.865	.95	.97
		N	318.31	412.89	31	100	189	354.8	2636
	Responsability	α	.903	.043	.79	.87	.915	.94	.96
		N	340.22	437.68	31	100	206.5	371	2636
	Unacceptable Thoughts	α	.911	.033	.83	.88	.92	.94	.96
		N	338.22	435.19	31	99	205	391	2636
	Symmetry	α	.913	.028	.85	.89	.92	.93	.96
		N	342.9	441.97	31	100	205	372	2636
PIR	Whole Scale	α	.92	.032	.83	.908	.925	.94	.96
		N	371.13	606.56	39	135.8	210	317.5	2976
	Impulses	α	.787	.059	.67	.76	.79	.84	.87
		N	461.65	704.56	93	150	223	360	2976
	Washing	α	.874	.061	.76	.83	.88	.92	.96
		N	459.47	705.79	58	150	222	360	2976
	Checking	α	.874	.029	.82	.858	.875	.89	.92
		N	484.88	720.96	120	192	243	367.8	2976
	Rumiation	α	.861	.044	.76	.83	.87	.88	.93
		N	461.59	704.55	93	150	224	360	2976
	Precision	α	.712	.085	.58	.658	.71	.778	.83
		N	484.94	720.92	120	192	244	367.5	2976
FOCI	Symptom	α	.819	.029	.75	.8	.83	.84	.86
		N	235.06	264.85	18	52	101	352	986
	Severity	α	.874	.051	.72	.86	.88	.91	.92
		N	356.6	329.68	47	87.5	352	437.5	1224
CAPS	PSP	α	.829	.053	.68	.8	.84	.86	.92
		N	399.79	448.76	37	82.3	236.5	582	1815
	PAO	α	.825	.068	.6	.8	.84	.87	.94
		N	458.26	548.74	37	90.5	253	569	2142
	Whole Scale	α	.827	.061	.6	.8	.84	.87	.94
		N	430	501.67	37	86	246	578	2142

Note: α = average alpha coefficient. N = sample size.

4.3.2 Comparing the procedures to obtain the average alpha coefficient

Overall, in terms of the average coefficient results, the 5% limit was not exceeded in any of the conditions compared. Calculating an average discrepancy index between the different scales in absolute values, the highest discrepancy percentage was 0.659%, comparing the conventional model with RVE applying Bonett’s transformation. Disregarding the results of the RVE model, the largest discrepancy was found within the conventional model when comparing the average alpha with and without transformation of the coefficients (0.553%). The lowest discrepancy index was 0.152%, comparing the homoscedastic and heteroscedastic multilevel models with transformed coefficients.

Conventional vs Multilevel model

Table 8 shows the results of the comparison between the conventional RG meta-analysis model (separating each subscale into independent meta-analyses) with the multilevel model that integrates all subscales within the same analysis establishing a hierarchical order (the whole-scale score has been considered as another subscale). Both models have been compared considering whether the primary coefficients were used raw -untransformed- or transformed using Bonett's formula -transformed-.

The discrepancy index never exceeded the 5% boundary. The highest value was found in the PI-R's *precision* subscale, when comparing the conventional model with the multilevel model, both with the coefficients untransformed (1.664%); while the lowest, also in absolute terms, was found on several occasions (0%). The highest absolute mean was found in the comparison between the conventional model and the RVE model when the coefficients were transformed (0.659%); while the lowest absolute mean was found in the comparison between conventional and multilevel model with untransformed coefficients (0.292%).

Table 8.

Discrepancy indices obtained comparing the multilevel model vs the conventional model.

Reference value: Conventional model		**Conventional vs Multilevel**				
		Untransformed			**Transformed**	
		k	**Multilevel**	**RVE**	**Multilevel**	**RVE**
DOCS	**Whole Scale**	72	.22	**0**	-.43	.22
	Contamination	58	.57	.68	-1.23	1.24
	Responsibility	50	**0**	.44	-.55	.77
	Unacceptable Thoughts	51	**0**	.11	-.33	.33
	Symmetry	49	-.111	.33	-.33	.55
PIR	**Whole Scale**	24	-.11	-.65	**0**	-.43
	Impulses	17	-.13	.76	-.5	1.77
	Washing	17	.23	.46	-.34	.34
	Checking	16	-.34	.34	-.34	.57
	Rumination	17	**0**	-1.04	-.23	-.92
	Precision	16	**1.66**	1.25	-.96	1.53
FOCI	**Symptom**	17	-.12	-.48	.37	-.61
	Severity	15	.11	-.11	.23	-.45
CAPS	**SPP**	58	.24	.24	.24	.12
	SOP	48	.49	.61	.36	.24
	Whole Scale	14	.35	-.35	.12	-.46
Average (absolute values)			**.292**	**.49**	**.409**	**.659**

Note: k = number of primary studies; *ML* = Multilevel Model; *RVE* = Robust Variance Estimator; *SPP* = Socially Prescribed Perfectionism; *SOP* = Self-Oriented Perfectionism

Conventional model vs Multilevel models (homo- and heteroscedastic models)

Table 9 shows the results of the comparison between the conventional meta-analysis model and two multilevel models: the homoscedastic and heteroscedastic model. The homoscedastic model assumes that the residuals error variance is homogeneously distributed across all subscales, whereas the heteroscedastic model calculates its own residuals error variance in each of the subscales (Hox, 2010). The three models have been compared considering whether the primary coefficients were used raw -untransformed- or transformed using Bonett's formula -transformed-.

The discrepancy index never exceeded the 5% boundary. The highest value was found in the PI-R's *precision* subscale, when comparing the conventional model with the homoscedastic multilevel model, both with the coefficients untransformed (1.664%); while the lowest, also in absolute terms, was found on several occasions (0). The highest absolute mean was found in the comparison between the conventional model and heteroscedastic multilevel model when the coefficients were transformed (0.313%); while the lowest absolute mean was found in the comparison between conventional model and homoscedastic multilevel model with transformed coefficients (0.193%).

Table 9.

Discrepancy indices obtained comparing the conventional model and the two different multilevel models (homo- and heteroscedastic model)

Reference value: Conventional model			**Conventional model vs Multilevel models**			
			Homoscedastic		**Heteroscedastic**	
Scales	**Subscales**	***k***	**UT**	**Bonett**	**UT**	**Bonett**
DOCS	**Whole Scale**	72	.22	**0**	-.11	-.32
	Contamination	58	.57	**0**	.68	.45
	Responsibility	50	**0**	**0**	.11	.11
	Unacceptable Thoughts	51	**0**	**0**	.11	.11
	Symmetry	49	-.11	**0**	**0**	**0**
PIR	**Whole Scale**	24	-.11	**0**	**0**	**0**
	Impulses	17	-.13	-.5	**0**	-.25
	Washing	17	.23	-.34	.11	-.79
	Checking	16	-.34	-.34	-.23	-.34
	Rumination	17	**0**	-.23	.12	-.35
	Precision	16	1.66	-.96	.83	-1.1

FOCI	**Symptom**	17	-.12	**0**	**0**	.24
	Severity	15	.11	**0**	**0**	**0**
CAPS	**SPP**	58	.24	.24	.24	.24
	SOP	48	.49	.36	.49	.49
	Whole Scale	14	.35	.12	.23	.23
Average (absolute values)			**.292**	**.193**	**.204**	**.313**

Note: k = number of primary studies*; UT* = Untransformed coefficients*; Bonett* = Bonett's transform coefficients; *SPP* = Socially Prescribed Perfectionism; *SOP* = Self-Oriented Perfectionism

Homoscedastic vs Heteroscedastic model

Table 10 shows the results of the comparison between two multilevel models, homo- and heteroscedastic models. Both models have been compared considering whether the primary coefficients were used raw -untransformed- or transformed using Bonett's formula -transformed-.

The discrepancy index never exceeded the 5% boundary. The highest value, in absolute terms, was found in the PI-R's *precision* subscale, when comparing the homoscedastic model with the heteroscedastic multilevel model, both with the coefficients untransformed (0.82%); while the lowest, also in absolute terms, was found on several occasions (0). The highest absolute mean was found in the comparison between the homo- and heteroscedastic models when the coefficients weren't transformed (0.157%), although the results are virtually identical when the transformation was applied (0.152%).

Table 10.

Discrepancy indices obtained comparing the two multilevel models: homoscedastic vs heteroscedastic model

Reference value:			**Homo- vs Heteroscedastic model**	
Homoscedastic model			**Heteroscedastic**	
Scales	**Subscales**	***k***	**Untransformed**	**Transformed**

DOCS	**Whole Scale**	72	-.32	-.32
	Contamination	58	.11	.45
	Responsibility	50	.11	.11
	Unacceptable Thoughts	51	.11	.11
	Symmetry	49	.11	**0**
PIR	**Whole Scale**	24	.11	**0**
	Impulses	17	.13	.25
	Washing	17	-.11	-.45
	Checking	16	.11	**0**
	Rumination	17	.12	-.12
	Precision	16	**-.82**	-.14
FOCI	**Symptom**	17	.12	.24
	Severity	15	-.11	**0**
CAPS	**SPP**	58	**0**	**0**
	SOP	48	**0**	.12
	Whole Scale	14	-.12	.11
Average (absolute values)			**.157**	**.152**

Note: k = number of primary studies; *SPP* = Socially Prescribed Perfectionism; *SOP* = Self-Oriented Perfectionism

Untransformed vs Transformed coefficients

Table 11 shows the results of the comparison within the two models (multilevel and conventional) between applying or not a transformation of the coefficients. Bonett's formula has been used because it has been found to work suitable in normalising the distribution and stabilising its variances (Badenes-Ribera et al., 2023; Sánchez-Meca et al., 2012).

The discrepancy index never exceeded the 5% boundary. The highest value, in absolute terms, was found in the DOCS' *contamination* subscale, when comparing the application of the Bonett's transformation within the conventional model (1.930%); while the lowest, also in absolute terms, was found on several occasions (0). The highest absolute mean was found in the comparison within the conventional model (0.553%); the lowest absolute mean was found within the comparison of the multilevel model (0.318%).

Table 11.

Discrepancy indices obtained comparing the untransformed average coefficient and the Bonett's transformation.

Reference value: *Untransformed coefficients*			**Untransformed vs Transformed**		
		k	**Conventional**	**Multilevel**	**RVE**
DOCS	**Whole Scale**	72	.32	-.32	.11
	Contamination	58	**1.93**	.11	1.24
	Responsibility	50	.66	.11	.44
	Unacceptable Thoughts	51	.33	**0**	.22
	Symmetry	49	.33	.11	.22
PIR	**Whole Scale**	24	.22	.33	.44
	Impulses	17	.25	-.13	.75
	Washing	17	1.48	.91	1.02
	Checking	16	.11	.11	**0**
	Rumination	17	.35	.12	.23
	Precision	16	.83	-1.77	.14
FOCI	**Symptom**	17	-.48	**0**	-.24
	Severity	15	-.11	**0**	-.23
CAPS	**SPP**	58	.24	.24	.36
	SOP	48	.61	.49	.61
	Whole Scale	14	.58	.34	.58
Average (absolute values)			**.553**	**.318**	**.426**

Note: k = number of primary studies*; RVE* = Robust Variance Estimator; *SPP* = Socially Prescribed Perfectionism; *SOP* = Self-Oriented Perfectionism

4.3.3 Comparing the procedures to obtain the confidence intervals of the average alpha coefficient

Contrary to what happened when comparing the different statistical methods for calculating the average coefficient, when applying these methods to the calculation of the confidence interval, we found that the 5% threshold was exceeded on most occasions. Calculating an average discrepancy index between the different scales in absolute values, the largest discrepancy percentage was 32.42%, comparing the conventional model with homoscedastic multilevel model applying Knapp-Hurtung's with raw coefficients. The lowest discrepancy index was 0.975%, comparing the standard method to calculate confidence intervals with Knapp-Hartung's method within the multilevel model with raw coefficients. That comparison in the method for calculating confidence intervals was the comparison with the least difference between conditions. The largest differences were found in Table 13 when comparing the conventional and multilevel models in their two versions (homo- and heteroscedastic).

Conventional vs Multilevel model

Table 12 provides the results of comparing the conventional RG meta-analysis model with the multilevel model when calculating confidence intervals around the average coefficient. Both models have been compared considering whether the primary coefficients were transformed by Bonett -transformed- or used raw -untransformed-; and whether the intervals were calculated according to the standard procedure -standard- or by applying the method proposed by Hartung and Knapp -Knapp-Hartung-.

The discrepancy index exceeded the 5% threshold in most cases. The highest value, in absolute terms, was found in the whole score of the PI-R scale, when comparing the conventional model with the RVE model, both with untransformed coefficients

(100%). Disregarding the RVE model, the highest discrepancy index was observed in FOCI's *symptom* subscale when the coefficients were transformed and the Knapp-Hartung formula was applied (91.67%). The lowest discrepancy index observed in absolute terms was found on four occasions (DOCS: *responsibility*; PI-R: *washing*; CAPS: *Self-Oriented Perfectionism -SOP-*), always in the condition in which the coefficients were transformed (0). The highest absolute mean was found in the comparison between the conventional model and the multilevel model when the raw coefficients were used and the Knapp-Hartung formula was applied (32.04%); the lowest absolute mean was found in the comparison between the conventional model and the RVE model when the coefficients were transformed (25.18%). The remaining comparisons showed very similar absolute means (26.56%; 26.17%; 26.71%).

Table 12.

Discrepancy indices obtained by comparing the conventional model and the multilevel model to calculate confidence intervals.

Reference value:			**Conventional vs Multilevel**					
Conventional Model			**Knapp-Hartung**		**Standard**		**RVE**	
Scales	**Subscales**	**k**	**UT**	**Bonett**	**UT**	**Bonett**	**UT**	**Bonett**
DOCS	**Whole Scale**	72	72.73	25	80	36.36	27.27	25
	Contamination	58	-41.67	-35.14	-38.24	-33.33	33.33	45.95
	Responsibility	50	-8.33	**0**	-4.35	**0**	-4.17	-4.55
	Unacceptable Thoughts	51	22.22	23.53	29.41	23.53	5.56	17.65
	Symmetry	49	37.5	46.67	37.5	40	6.25	6.67
PIR	**Whole Scale**	24	50	-4.17	71.43	4.55	**100**	83.33
	Impulses	17	-20.69	31.03	-13.21	37.04	-10.35	-6.9
	Washing	17	-30.16	-34.92	-22.81	-29.31	-12.7	**0**
	Checking	16	45.16	40.63	57.14	50	16.13	12.5
	Rumination	17	18.42	6.67	13.16	12.19	97.37	77.78

	Precision	16	-43.82	13.33	-37.5	21.69	16.85	20
FOCI	**Symptom**	17	43.48	**91.67**	33.33	72	26.09	33.33
	Severity	15	-33.33	-31.82	-17.65	-25.64	7.14	22.73
CAPS	**SPP**	58	7.69	16	16.67	16	11.54	20
	SOP	48	-18.42	-8.11	-13.89	-8.11	-2.63	**0**
	Whole Scale	14	19.05	-16.33	19.05	-8.89	50	26.53
Average (absolute values)			**32.042**	**26.562**	**31.583**	**26.165**	**26.711**	**25.182**

Note: UT = Untransformed*; RVE* = Robust Variance Estimator*; k* = number of primary studies; *SPP* = Socially Prescribed Perfectionism; *SOP* = Self-Oriented Perfectionism

Conventional vs Multilevel models (homo- and heteroscedastic models)

Table 13 presents the results of comparing the conventional RG meta-analysis model with the two multilevel models (homo- and heteroscedastic model) when calculating confidence intervals around the average coefficient. All models have been compared considering whether the primary coefficients were transformed by Bonett -transformed- or used raw -untransformed-; and whether the intervals were calculated according to the standard procedure -standard- or by applying Knapp and Hartung’s method -Knapp-Hartung-.

The discrepancy index exceeded the 5% threshold in most cases. The highest value, in absolute terms, was found in the FOCI’s *symptom* subscale when comparing the conventional model with the homoscedastic model, both with transformed coefficients and applying Knapp-Hartung’s method (91.67%). The lowest discrepancy index observed in absolute terms was found in different occasions (0). The highest absolute mean was found in the comparison between the conventional model and the homoscedastic multilevel model when the raw coefficients were used and the Knapp-Hartung formula was applied (32.04%); the lowest absolute mean was found in the comparison between the conventional model and the heteroscedastic model when the coefficients were transformed and applied Knapp-Hartung’s method (8.23%). It should be noted that in the conditions where the conventional model was compared with the heteroscedastic model, the results of the absolute mean of all comparisons resulted in values very close to 10% discrepancy.

Table 13.

Discrepancy indices obtained by comparing the conventional model and the two multilevel models to calculate confidence intervals.

Reference value: Conventional model			**Conventional vs Multilevel**							
			Knapp-Hartung				**Standard**			
			Homoscedastic		**Heteroscedastic**		**Homoscedastic**		**Heteroscedastic**	
Scales	**Subscales**	**k**	**UT**	**Bonett**	**UT**	**Bonett**	**UT**	**Bonett**	**UT**	**Bonett**
DOCS	**Whole Scale**	72	72.73	25	18.18	-8.33	80	36.364	30	**0**
	Contamination	58	-41.67	-35.14	-5.56	-5.41	-38.24	-33.33	**0**	-2.78
	Responsibility	50	-8.33	**0**	**0**	4.55	-4.35	**0**	4.35	4.545
	Unacceptable Thoughts	51	22.22	23.53	11.11	11.77	29.41	23.53	17.65	11.765
	Symmetry	49	37.5	46.67	25	26.67	37.5	40	25	26.667
PIR	**Whole Scale**	24	50	-4.17	12.5	-4.17	71.43	4.55	28.57	4.545
	Impulses	17	-20.69	31.03	-24.14	-10.35	-13.21	37.047	-18.87	-3.704
	Washing	17	-30.16	-34.92	-6.35	3.18	-22.81	-29.31	1.75	12.069
	Checking	16	45.16	40.63	25.81	18.75	57.14	50	39.29	26.667
	Rumination	17	18.42	6.67	2.63	**0**	13.16	12.2	**0**	9.756
	Precision	16	-43.82	13.33	-13.48	7.78	-37.5	21.69	-5.	16.867
FOCI	**Symptom**	17	43.48	**91.67**	8.7	8.33	33.33	72	**0**	**0**
	Severity	15	-33.33	-31.82	-9.52	**0**	-17.65	-25.64	5.88	7.692
CAPS	**SPP**	58	7.69	16	**0**	4	16.67	16	4.17	**0**

SOP	48	-18.42	-8.11	**0**	**0**	-13.89	-8.11	**0**	**0**
Whole Scale	14	19.05	-16.33	2.38	-18.37	19.05	-8.89	-2.38	-11.111
Average (absolute values)		**32.042**	**26.562**	**10.335**	**8.227**	**31.583**	**26.165**	**11.432**	**8.635**

Note: UT = untransformed*; k* = number of primary studies; *SPP* = Socially Prescribed Perfectionism; *SOP* = Self-Oriented Perfectionism

Homoscedastic vs Heteroscedastic model

Table 14 contains the results of comparing the two multilevel models (homo- and heteroscedastic) with each other when calculating confidence intervals around the average coefficient. Both models have been compared considering whether the primary coefficients were transformed by Bonett -transformed- or used raw -untransformed-; and whether the intervals were calculated according to the standard procedure -standard- or by applying Knapp and Hartung's method -Knapp-Hartung-.

The discrepancy index exceeded the 5% boundary on a large number of occasions. The highest value, in absolute terms, was found in DOCS' *contamination* subscale when the homoscedastic model was compared with the heteroscedastic model, with raw coefficients and applying both methods of estimating the confidence intervals (61.91%).

The lowest discrepancy index observed in absolute terms was found in the total score of the PI-R scale (0), when the coefficients were transformed and with the two interval estimation methods. The highest absolute mean was found in the comparison between the homoscedastic model and the heteroscedastic model when the raw coefficients were used and the Knapp-Hartung formula was applied (23.03%); the lowest absolute mean was found in the comparison between the homo- and the heteroscedastic model when the coefficients were transformed and Knapp-Hartung was not applied (19.86%).

Table 14.

Discrepancy indices obtained by comparing the homo- and heteroscedastic model to calculate confidence intervals.

Reference value: Homoscedastic model			**Homo- vs Heteroscedastic**			
			Knapp-Hartung's		**Standard**	
Scales	**Subscales**	**k**	**UT**	**Bonett**	**UT**	**Bonett**
DOCS	**Whole Scale**	72	-31.58	-26.67	-27.78	-26.67
	Contamination	58	**61.91**	45.83	**61.91**	45.83
	Responsibility	50	9.09	4.55	9.09	4.55
	Unacceptable Thoughts	51	-9.09	-9.52	-9.09	-9.52
	Symmetry	49	-9.09	-13.64	-9.09	-9.52
PIR	**Whole Scale**	24	-25	**0**	-25	**0**
	Impulses	17	-4.35	-31.58	-6.52	-29.73
	Washing	17	34.09	58.54	31.82	58.54
	Checking	16	-13.33	-15.56	-11.36	-15.56
	Rumination	17	-13.33	-6.25	-11.63	-2.17
	Precision	16	54	-4.90	52	-3.96
FOCI	**Symptom**	17	-24.24	-43.48	-25	-41.86
	Severity	15	35.71	46.67	28.57	44.83
CAPS	**SPP**	58	-7.14	-10.35	-10.71	-13.79
	SOP	48	22.58	8.82	16.13	8.82
	Whole Scale	14	-14	-2.44	-18	-2.44
Average (absolute values)			**23.034**	**20.549**	**22.106**	**19.862**

Note: UT = untransformed*; k* = number of primary studies; *SPP* = Socially Prescribed Perfectionism; *SOP* = Self-Oriented Perfectionism

Untransformed vs transformed coefficients

Table 15 collects the results of comparing within each model -conventional and multilevel-, the influence of transforming the primary coefficients when constructing confidence intervals around the average coefficient. Both models have been compared considering whether the intervals were calculated according to the standard procedure or by applying the method of Knapp and Hartung.

The discrepancy index exceeded the 5% boundary on part of the cases. The highest value, in absolute terms, was found in PI-R's *precision* subscale when the influence of transforming was compared within multilevel model applying Knapp-Hartung's method (104%). The lowest discrepancy index observed in absolute terms was found in several cases (0). The highest absolute mean was found in the comparison within the multilevel model applying Knapp-Hartung (21.03%); the lowest absolute mean was found within the conventional model applying the standard method (5.41%). It should be noted that, except for the multilevel model, the discrepancy index was close to 5% in all comparisons.

Table 15.

Discrepancy indices obtained by comparing within the conventional model and the multilevel model, the application of a transformation of the coefficients to calculate confidence intervals.

Reference value:			Untransformed vs Transformed				
Untransformed coefficients			**Knapp-Hartung**		**Standard**		**RVE**
Scales	**Subscales**	***k***	**Conventional**	**Multilevel**	**Conventional**	**Multilevel**	
DOCS	**Whole Scale**	72	9.09	-21.05	10	-16.67	7.14
	Contamination	58	2.78	14.29	5.88	14.29	12.5
	Responsibility	50	-8.33	**0**	-4.35	**0**	-8.7
	Unacceptable Thoughts	51	-5.56	-4.55	**0**	-4.55	5.26
	Symmetry	49	-6.25	**0**	-6.25	-4.55	-5.88
PIR	**Whole Scale**	24	**0**	-36.11	4.76	-36.11	-8.33
	Impulses	17	**0**	65.22	1.89	60.87	3.85
	Washing	17	**0**	-6.82	1.75	-6.82	14.55
	Checking	16	3.23	**0**	7.14	2.27	**0**

	Rumination	17	18.42	6.67	7.9	6.98	6.67
	Precision	16	1.12	**104**	3.75	102	3.85
FOCI	**Symptom**	17	4.35	39.39	4.17	34.38	10.35
	Severity	15	4.76	7.14	14.71	3.57	20
CAPS	**SPP**	58	-3.85	3.57	4.17	3.57	3.45
	SOP	48	-2.63	9.68	2.78	9.68	**0**
	Whole Scale	14	16.67	-18	7.14	-18	-1.59
Average (absolute values)			**5.440**	**21.030**	**5.414**	**20.268**	**7.006**

Note: k = number of primary studies; *RVE* = Robust Variance Estimator; *SPP* = Socially Prescribed Perfectionism; *SOP* = Self-Oriented Perfectionism

Knapp-Hartung's method vs standard method

Table 16 presents the results of comparing within each model -conventional and multilevel- the influence of calculating confidence intervals using the Standard method or applying the Knapp-Hartung's method. Both models have been compared considering whether the primary coefficients were transformed by Bonett's -transformed- or used raw -untransformed-.

The discrepancy index exceeded the 5% boundary in very few cases. The highest value, in absolute terms, was found in FOCI's *severity* subscale within the conventional model using the raw coefficients (23.53%). The lowest discrepancy index observed in absolute terms was found in most cases (0). The highest absolute mean was found in the comparison within the conventional model using raw coefficients (7.74%); the lowest absolute mean was found within the multilevel model when the coefficients weren't transformed (0.98%). It should be noted that the discrepancy index was close to 5% in all comparisons.

Table 16.

Discrepancy indices obtained by comparing within the conventional model and the multilevel model, the application of Knapp-Hartung's method to calculate confidence intervals

Reference value: *Standard Method*			Standard vs Knapp-Hartung's method			
			Untransformed		Transformed	
Scales	**Subscales**	**k**	**Conventional**	**Multilevel**	**Conventional**	**Multilevel**
DOCS	**Whole Scale**	72	10	5.56	9.09	**0**
	Contamination	58	5.88	**0**	2.78	**0**
	Responsibility	50	4.35	**0**	**0**	**0**
	Unacceptable Thoughts	51	5.88	**0**	**0**	**0**
	Symmetry	49	**0**	**0**	**0**	4.76
PIR	**Whole Scale**	24	14.29	**0**	9.09	**0**
	Impulses	17	9.43	**0**	7.41	2.7
	Washing	17	10.53	**0**	8.62	**0**
	Checking	16	10.71	2.27	6.67	**0**
	Rumination	17	**0**	4.65	9.76	4.35
	Precision	16	11.25	**0**	8.43	.99
FOCI	**Symptom**	17	-4.17	3.13	-4	6.98
	Severity	15	**23.53**	**0**	12.82	3.45
CAPS	**SPP**	58	8.33	**0**	**0**	**0**
	SOP	48	5.56	**0**	**0**	**0**
	Whole Scale	14	**0**	**0**	8.89	**0**
Average (absolute values)			**7.744**	**0.975**	**5.472**	**1.452**

Note: k = number of primary studies; *SPP* = Socially Prescribed Perfectionism; *SOP* = Self-Oriented Perfectionism

4.4 Discussion

In this study we have compared different analytical strategies to carry out a meta-analysis of reliability generalization when the scale has several subscales. As mentioned above, this type of scales can lead to dependence between scores because the same participants complete all the subscales and, at the same time, these subscales represent different dimensions of the same construct. As a first approach to this type of comparison studies between techniques and models, we thought it congruent to apply it to real data from meta-analyses already published in impact journals. We believe that this work represents a very novel study both in the field of multilevel analysis and in the field of RG meta-analysis, since so far there is no study that applies this type of 3-level structure to psychometric scales with more than one subscale.

Evaluating the results in the calculation of the average coefficient, we have observed that there are practically no differences between the models applied. The numerical results are around 0-1% (in absolute terms) and in no circumstance has the limit of 5% been exceeded. We found that the largest discrepant value was 1.93%, which occurred in the condition where the transformations of the coefficients within each model were compared. On this case, this percentage corresponds to comparing the average coefficient of the *Contamination* subscale of the DOCS in the conventional model, with the reference value being the coefficient without transformation and the compared value being the transformed coefficient. The minimum discrepancy value was 0%, which was repeated many times throughout all comparison conditions. Nor was any pattern observed according to the number of subscales or the number of primary coefficients of the chosen scales.

On the other hand, the results concerning confidence width are more complex. First, we found that the discrepancy indices calculated here were systematically higher than 5%. This result indicated that the main differences in following one or the other analytical strategy are to be found when calculating the confidence intervals of the average coefficient. Of the five comparisons that were carried out, the comparison with the lowest values was the one that tested the method of estimating the intervals within each model (Table 16). That is, the comparison that controlled whether the estimation had been performed by the standard method or whether the Knapp and Hartung’s improved formula had been applied. The results were to be expected: the percentage discrepancy of more than 5% was found in the conventional method (7.74% and 5.47%) and not in the multilevel model (0.98% and 1.45%). Moreover, a 2% decrease in the discrepancy was observed when the coefficients were transformed. The highest value was found in the *Severity* subscale of the FOCI scale (23.53%) when the coefficients were not transformed within the conventional model.

A relevant result was found in the comparison between transforming or not transforming the coefficients within each model (Table 15). While, in the conventional model, the results were around 5%, the multilevel model showed average variations of 20%. The highest discrepancy index appeared in the *Precision* subscale of the PI-R scale (104%). The RVE model seemed to represent an intermediate point of discrepancy between the two models, although it tended more towards the conventional model, with an average discrepancy rate of 7%. In other words, it seemed that the decision between transforming the coefficients or not could imply more different results in the multilevel model than in the conventional model. What has also not been observed is that these results are not very different regarding whether Knapp-Hartung’s method was applied or not (21.03% vs. 20.27% in the multilevel model; 5.44% vs. 5.41% in the conventional model).

Another consistent result was found in Table 13, when comparing the conventional model with the two forms of the multilevel model (homo- and heteroscedastic). While, when comparing the homoscedastic model with the conventional model the results reached 32%, when compared with the heteroscedastic model the results were found to be between 8 and 11%. This is explained by the fact that, when separating each scale into independent analyses, these results have more similarities with the heteroscedastic model, which calculates the residuals error variance for each component of the model, than with the homoscedastic model, since that model estimates the average residuals error variance of all components. Here again, we did not observe large distances between the application or non-application of Knapp-Hartung's method (32% vs. 31.5% in homo- model; 10% vs. 11% in hetero- model; 26.5% vs. 26% in homo- model with transformation; 8% vs. 8.6% in hetero- model with transformation). Where we did find a small variation was when we considered whether or not the coefficients were transformed. We found that the homoscedastic model went from 32% to 26.5% discrepancy if we transformed the coefficients; while in the heteroscedastic model, the discrepancy went from 10% to 8%. When comparing the two multilevel models, the results are completely expected: around 20% discrepancy in all conditions. Finally, the relationship between the conventional model and the RVE model presented in Table 12 showed very similar results to those found between the conventional model and the homoscedastic model (~26% discrepancy), a result that was expected and consistent with the model.

Regarding the relationship between the number of observations and the number of subscales, there did not seem to be a clear pattern of trend in any of the five comparisons. To study this relationship in depth, it would be interesting to carry out a simulation study to better understand how each of the analysis strategies we have tested works and what influence the conditions have on each of the strategies.

4.4.1. Conventional model or multilevel model?

So which model is the most appropriate? An empirical study of this kind could never draw conclusions about the advantages of one model or the other, but what we can highlight is the trend we have found in the results. Regarding the estimation of the average alpha, we have not found any variation that depends on the specific analytical strategy, so we are going to focus on highlighting only the conclusions drawn from the results in terms of the confidential width of the average alpha.

First, when applying the multilevel model, we recommend transforming the coefficients. As we already mentioned, in this study we have found very discrepant results between not transforming and transforming the coefficients (~20%). Great part of the literature recommends transforming in all circumstances (Sánchez-Meca et al., 2012), although the most recommendable transformations -for internal consistency coefficients- are those proposed by Hakstian and Whalen (1976) and Bonett (2002). As we discussed in Chapter 2, these transformations normalize the distribution of the coefficients and, in addition, Bonett's also stabilizes their variances.

Regarding the application of Knapp and Hartung's method, we also found favourable results for the application of this method in the previous literature. However, in our results the differences between applying it or not have been limited, exceeding the 5% boundary only within the conventional model (~6%). In other words, we believe that it would be advisable to apply it when conducting a meta-analysis in the conventional model, but not as indispensable within the multilevel model.

Finally, the great question would be whether we would choose to apply the conventional model or the multilevel model when carrying out a meta-analysis of reliability generalization. Based on the results we have obtained, we are able to say that the differences between one and the other are notable (with any of the multilevel models, the differences exceed 5%), but we cannot highlight the advantages of one model over the other. From a theoretical point of view, an RG study with scales that have several subscales or with scales that, in the different primary studies, have been administered to different groups within the same study, the application of the multilevel model would be more accurate. This is because the inclusion of these two assumptions would yield dependency between scores. In fact, given the assumptions, the question would be whether to choose between the homoscedastic model or the heteroscedastic model, since we also found discrepant results between the two models.

4.4.2. Limitations and Future Research

The main limitation of this study is that the conclusions are reduced to detecting only whether there are notable differences between the analysis strategies, but we cannot determine which strategy is the most appropriate or the most accurate on each circumstance. In order to be able to go deeper into this study and have empirical results that allow us to conclude on the suitability of each model for each specific case, it would be interesting to carry out a simulation study. On the other hand, from one of our hypotheses on the influence of the number of observations and the number of scales, we have not been able to draw clear conclusions that we could control and determine through a simulation. It would also be interesting to include in that study how the moderating variables that have traditionally been found to influence the data in meta-analyses of reliability generalization -such as standard deviation of the scores- work in each of the models: conventional and multilevel.

4.5 Conclusion

In this research we have evaluated the discrepancies in the results of meta-analysing four different scales composed of different subscales by means of different analysis strategies. We found that the differences appear mainly when estimating the confidence intervals of the average reliability coefficients; when these coefficients were estimated, no notable differences appeared between any of the different strategies implemented. The main results indicate that when estimating the confidential width discrepancies appear between the conventional model and the two multilevel models, being the differences between the conventional and the homoscedastic multilevel model and between the two multilevel models and each other more remarkable. Also, the results suggest that when the multilevel model is applied, the transformation of the coefficients is substantially more important than in the conventional model, as the discrepancy was very high between the two conditions. The next step in this research is to develop a simulation that will allow us to obtain information on which model is more appropriate when conducting a reliability generalization meta-analysis under certain circumstances.

Chapter 5

Conclusions

Reliability is a psychometric property that refers to the replicability of the scores on a measuring instrument. As the definition suggests, this property is not inherent to the instrument, but to the scores, so it is essential to apply statistical tools that allow us to generalize the results obtained from the different applications of an instrument to the instrument itself. To date, the best tool for synthesising quantitative evidence is meta-analysis. A key aspect that makes meta-analysis stand out is that, on most occasions, the primary studies may yield different and even contradictory results; however, meta-analysis allows for grouping all these results together and obtaining a more accurate value that integrates all the information. This methodology can also be applied to reliability coefficients. Vacha-Haase (1998) developed this concept and called it Reliability Generalization Meta-Analysis (RG).

As we have seen throughout this dissertation, there is no single protocol for application, but rather it is the responsibility of the meta-analyst to decide which analytical strategy is most appropriate for the type of study being carried out. Because of this variability in decision-making, this dissertation has elaborated two comparative studies between different statistical procedures for the computation of an RG meta-analysis (Chapter 2 and Chapter 4). The first study (Chapter 2) compared the different strategies most frequently used in this field in a conventional way, i.e., without considering any kind of dependency relationship. However, the third study (Chapter 4) compared this conventional procedure with a procedure that takes into account these dependency networks within studies and scales by applying different multilevel models.

The main conclusion we found in both Chapter 2 and Chapter 4 is that, as expected, there are statistically significant differences between applying one procedure and the other.

Beginning with the first study and the first comparisons between analytical strategies, 138 databases of RG meta-analyses on psychological scales and subscales were collected. The statistical procedures most frequently used for such meta-analyses were applied to these databases: on the one hand, we selected three transformations that are regularly applied to the reliability coefficients and compared them also with the raw - untransformed - coefficients; on the other hand, we also took into account the assumed statistical model and the weighting method. Due to the theoretical nature of some strategies, we compared 13 procedures for the computation of the average reliability coefficient and 18 procedures for the computation of the confidence interval. We also took into account whether coefficient transformations influenced the distribution of coefficients and the estimation of different heterogeneity indices, such as the I^2 index and the prediction intervals.

The first conclusion we can draw from this study is that, in numerical terms, the different procedures significantly affect the construction of the confidence interval of the average reliability coefficient, but not the calculation of the average reliability coefficient. Nevertheless, we did find statistically significant results at the level of the average coefficient when we looked at how the transformations influenced its distribution: all the proposed transformations improved the adjustment to normality, in some cases bringing it closer to a more platykurtic distribution, and also improved the degree of skewness of the distribution.

Regarding the construction of the confidence interval, it is the assumed statistical model and not the transformations that determine whether the confidential width will be wider or narrower. As expected, the OLS and REi models presented a wider width of the interval than the other models, the narrowest being the VC and FE models, the latter being the one with the narrowest interval of all. Furthermore, within the RE models, no differences were observed between the DL and REML τ^2 estimators.

In Chapter 4, comparisons were made considering any dependency relationships that may arise within the studies that comprise the meta-analysis. In this study we have compared the conventional method of calculating a meta-analysis with scales that are composed of several subscales (separating each one into an independent meta-analysis) and the multilevel model that integrates them all in the same meta-analysis. In addition, within this model we have also compared the homoscedastic and heteroscedastic multilevel model. Complementarily, comparisons with the RVE model, Bonett's transformation of the reliability coefficients and Hartung-Knapp's method for the construction of confidence intervals were included. All this was applied to 4 psychological scales that had been the subject of a published RG MA and differed from each other in terms of number of subscales (less or more than 4) and number of observations per subscale (less or more than 20 observations).

The main conclusion we draw from this study coincides with that obtained in Chapter 2: the numerically significant results were found when constructing the confidence interval, but not when computing the average coefficient. The main differences in terms of discrepancy index were found when comparing this conventional model with the homoscedastic multilevel model (approx. 32% with transformation, 26% without transformation), reducing as expected when compared with the heteroscedastic model (approx. 10% without transformation, 8% with transformation). On the other hand, Bonett's transformation of the coefficients showed a larger discrepancy within the multilevel model (approx. 21%) than within the conventional model (approx. 8%). The application of the Hartung-Knapp method did not seem to have any influence on the results of any of the models, nor was there a clear trend in the results taking into account the number of subscales or the number of observations.

Finally, Chapter 3 reports a reproducibility study of RG meta-analyses, including a review of their transparency and reporting practices. Due to the very large database collected in Chapter 2, this study was conducted in parallel and was intended to be a first contact between reproducibility studies and RG meta-analyses, something that had not been considered before. In addition to the databases collected in Chapter 2, this study compiled all the information reported by the studies on the estimation method used to calculate the meta-analysis and repeated the analyses following these procedures. To determine the extent to which the results were reproduced, a discrepancy index between the reported and reproduced results and the Pearson correlation was calculated.

The main conclusion that can be drawn from this study is that in all the variables that were tested (average reliability coefficient, confidence interval and heterogeneity indices I^2 and Q), the correlation between the reported and reproduced values was higher than .80 and statistically significant. Regarding the discrepancy index, not all variables responded equally: on the one hand, both the average coefficient and its confidence interval were reproduced more than 95% of the time, with only anecdotal cases where the discrepancy exceeded 10% variation. However, the results on heterogeneity did show greater variability, mainly in the *Q* statistic, where the discrepancy index showed values of less than 10% variation in 75% of the cases. One of the conclusions we draw from this is that, as the value of the *Q* statistic is a value that fluctuates between 0 and $+\infty$, it is more sensitive to large numerical variations. This leads to the discrepancy index also being higher. In this case, the correlation value complements the results of the discrepancy index by determining that the variations that occur do not affect breeding success as directly ($r_{xy} = .99$).

Finally, the last conclusion that we highlight in this chapter is the great loss of information that occurs due to the systematic lack of reporting in this type of work. Not only the absence of certain important results such as the confidence interval or a heterogeneity index, but also the loss of data in relation to the statistical procedures implemented or the lack of reporting of the database used to perform the meta-analysis. Almost 60% of the studies found in the search did not share the database. On the other hand, with respect to the variables that have been reproduced, of the total alpha coefficients collected, in 40% of the cases the confidence interval was not reported, and in more than 50% no heterogeneity index was reported.

In recent years, tools have been developed and implemented which make it easier to correct these practices, such as the use of reporting guidelines and free online repositories where the materials used in the studies, such as databases or programming codes, can be stored. It is now up to researchers to take responsibility for being more aware of the importance of sharing everything necessary for research to be reproducible and replicable. We should not forget that one of the fundamental pillars of science is the replicability of experimental results and that such replications only ensure that the conclusions are solid and stable.

In summary, and taking into account the results obtained in this thesis, the main recommendations are the following:

The researcher carrying out this type of meta-analysis must be clear about the nature of the data to be analyzed (presence of heterogeneity and relationships, or suspicions, of dependence). In addition, it is essential to establish the scope of its results (more or less generalisable). For an RG meta-analysis without the presence of heterogeneity or dependence relationships, which aims to generalize the results to studies with identical or very similar characteristics, it should choose to apply an FE model. If heterogeneity is present, then the assumed model must be a Varying Coefficients model.

On the other hand, if the results are intended to be generalized to a larger population of studies, and there is no suspicion of dependence networks, the random effects model in one of its forms (RE, REi and REn) should be the model of choice. In this case, it is important to note that random effects models must meet three fundamental assumptions: normality of the true reliability coefficient distribution, a stable estimate of the between-studies variance, and random sampling of studies from a larger population of primary studies.

In any of the above cases, if it is suspected that dependency relationships may exist within the meta-analysis, it would be advisable to perform it from a multilevel approach applying a homo- or heteroscedastic model.

These recommendations, coming from empirical studies with real data, do not determine which model works best or is the most appropriate according to a particular type of data. These recommendations are based on theoretical knowledge prior to this work and how it can be applied taking into account the results obtained (i.e., that there are differences between performing an RG meta-analysis following different statistical procedures).

A final recommendation from Chapter 3 is the use of reporting guidelines and online repositories to improve the transparency and reporting rates of this type of meta-analysis. Specifically, for RG meta-analyses, the REGEMA checklist (Sánchez-Meca et al., 2021) is designed to correctly report the necessary and fundamental information on this type of meta-analysis.

References

Abad, F. J., Olea-Díaz, J., Ponsoda-Gil, V., & García-García, C. (2011). *Medición en ciencias sociales y de la salud*. Síntesis.

Abramowitz, J. S., Deacon, B. J., Olatunji, B. O., Wheaton, M. G., Berman, N. C., Losardo, D., Timpano, K. R., McGrath, P. B., Riemann, B. C., Adams, T., Björgvinsson, T., Storch, E. A., & Hale, L. R. (2010). Assessment of obsessive-compulsive symptom dimensions: Development and evaluation of the Dimensional Obsessive-Compulsive Scale. *Psychological Assessment*, *22*(1), 180–198. https://doi.org/10.1037/a0018260

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

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, *26*(5), 527–546. https://doi.org/10.1037/met0000365

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology*, *12*(3), 154–174. https://doi.org/10.20982/tqmp.12.3.p154

Badenes-Ribera, L., Duro-García, C., López-Ibáñez, C., Martí-Vilar, M., & Sánchez-Meca, J. (2023). The Adult Prosocialness Behavior Scale: A reliability generalization meta-analysis. *International Journal of Behavioral Development*, *47*(1), 59–71. https://doi.org/10.1177/01650254221128280

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

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

Bonett, D. G. (2002). Sample Size Requirements for Testing and Estimating Coefficient Alpha. *Journal of Educational and Behavioral Statistics*, *27*(4), 335–340. https://doi.org/10.3102/10769986027004335

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

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

Borenstein, M., & Hedges, L. V. (2019). Effect Sizes for Meta-Analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine, *The Handbook of Research Synthesis and Meta-Analysis* (pp. 2018–2255). Russell Sage Foundation.

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

Botella, J., & Sánchez-Meca, J. (2015). *Meta-análisis en ciencias sociales y de la salud*. Síntesis.

Breidbord, J., & Croudace, T. J. (2013). Reliability Generalization for Childhood Autism Rating Scale. *Journal of Autism and Developmental Disorders*, *43*(12), 2855–2865. https://doi.org/10.1007/s10803-013-1832-9

Cohen, J. (1977). The Significance of a Product Moment. In *Statistical Power Analysis for the Behavioral Sciences* (pp. 75–108). Academic Press.

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

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/10.1007/bf02310555

Fernández-Castilla, B., Jamshidi, L., Declercq, L., Beretvas, S. N., Onghena, P., & Van Den Noortgate, W. (2020). The application of meta-analytic (multi-level) models with multiple random effects: A systematic review. *Behavior Research Methods*, *52*(5), 2031–2052. https://doi.org/10.3758/s13428-020-01373-9

Fernández-Castilla, B., Maes, M., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van Den Noortgate, W. (2019). A demonstration and evaluation of the use of cross-classified random-effects models for meta-analysis. *Behavior Research Methods*, *51*(3), 1286–1304. https://doi.org/10.3758/s13428-018-1063-2

Fisher, Z., & Tipton, E. (2015). *robumeta: An R-package for robust variance estimation in meta-analysis* (arXiv:1503.02220). arXiv. http://arxiv.org/abs/1503.02220

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

Flett, G. L., Hewitt, P. L., Besser, A., Su, C., Vaillancourt, T., Boucher, D., Munro, Y., Davidson, L. A., & Gale, O. (2016). The Child–Adolescent Perfectionism Scale: Development, Psychometric Properties, and Associations With Stress, Distress, and Psychiatric Symptoms. *Journal of Psychoeducational Assessment*, *34*(7), 634–652. https://doi.org/10.1177/0734282916651381

Glass, G. V. (1976). Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher*, *5*(10), 3–8. https://doi.org/10.3102/0013189X005010003

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

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

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, C., Kidwell, M. C., Mohr, A. H., Clayton, E., Yoon, E. J., Henry, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science, 5*, 1–18. http://dx.doi.org/10.1098/rsos.180448

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

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Henson, R. K., & Thompson, B. (2002). Characterizing Measurement Error in Scores Across Studies: Some Recommendations for Conducting “Reliability Generalization” Studies. *Measurement and Evaluation in Counseling and Development, 35*(2), 113–127. https://doi.org/10.1080/07481756.2002.12069054

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

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2. ed). Routledge, Taylor & Francis.

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

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

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

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

Laird, N. M., & Mosteller, F. (1990). Some Statistical Methods for Combining Experimental Results. *International Journal of Technology Assessment in Health Care*, *6*(1), 5–30. https://doi.org/10.1017/S0266462300008916

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

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

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *Annals of Internal Medicine*, *151*(4), 65–94. https://doi.org/10.7326/0003-4819-151-4-200908180-00136

López-Nicolás, R., Rubio-Aparicio, M., López-Ibáñez, C., & Sánchez-Meca, J. (2021). A Reliability Generalization Meta-analysis of the Dimensional Obsessive-

Compulsive Scale. *Psicothema*, *33.3*, 481–489. https://doi.org/10.7334/psicothema2020.455

López-Pina, J. A., Sánchez-Meca, J., López-López, J. A., Marín-Martínez, F., Núñez-Núñez, R. M., Rosa-Alcázar, A. I., Gómez-Conesa, A., & Ferrer-Requena, J. (2015). The Yale–Brown Obsessive Compulsive Scale: A Reliability Generalization Meta-Analysis. *Assessment*, *22*(5), 619–628. https://doi.org/10.1177/1073191114551954

Maassen, E., van Assen, M. A. L. M., Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLOS ONE*, *15*(5), e0233107. https://doi.org/10.1371/journal.pone.0233107

Maes, M., Van den Noortgate, W., & Goossens, L. (2015). A Reliability Generalization Study for a Multidimensional Loneliness Scale: The Loneliness and Aloneness Scale for Children and Adolescents. *European Journal of Psychological Assessment*, *31*(4), 294–301. https://doi.org/10.1027/1015-5759/a000237

Mason, C., Allam, R., & Brannick, M. T. (2007). How to Meta-Analyze Coefficient-of-Stability Estimates: Some Recommendations Based on Monte Carlo Studies. *Educational and Psychological Measurement*, *67*(5), 765–783. https://doi.org/10.1177/0013164407301532

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

Muñiz, J. (2018). *Introducción a la Psicometría*. Pirámide.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022).

Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Núñez-Núñez, R. M., Rubio-Aparicio, M., Marín-Martínez, F., Sánchez-Meca, J., López-Pina, J. A., & López-López, J. A. (2022). A Reliability Generalization Meta-Analysis of the Padua Inventory-Revised (PI-R). *International Journal of Clinical and Health Psychology*, *22*(1), 100277. https://doi.org/10.1016/j.ijchp.2021.100277

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

Pashler, H., & Wagenmakers, E. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Pedersen, T. (2022). *patchwork: The Composer of Plots.* (1.1.2.9000) [R]. https://patchwork.data-imaginist.com, https://github.com/thomasp85/patchwork.

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with Robust Variance Estimation: Expanding the Range of Working Models. *Prevention Science*, *23*(3), 425–438. https://doi.org/10.1007/s11121-021-01246-3

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

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling Multivariate Effect Sizes. *Psychological Bulletin*, *103*(1), 111–120.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes Meta-Analysis. *Journal of Educational Statistics*, *10*(2), 75–98. https://doi.org/10.3102/10769986010002075

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*(3), 306–322. https://doi.org/10.1037/1082-989X.11.3.306

Romano, J. L., Kromrey, J. D., & Hibbard, S. T. (2010). A Monte Carlo Study of Eight Confidence Interval Methods for Coefficient Alpha. *Educational and Psychological Measurement*, *70*(3), 376–393. https://doi.org/10.1177/0013164409355690

Rubio-Aparicio, M., Badenes-Ribera, L., Sánchez-Meca, J., Fabris, M. A., & Longobardi, C. (2020). A reliability generalization meta-analysis of self-report measures of muscle dysmorphia. *Clinical Psychology: Science and Practice*, *27*(1). https://doi.org/10.1111/cpsp.12303

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

Sánchez-Meca, J., López-López, J. A., & López-Pina, J. A. (2012). Some recommended statistical analytic practices when reliability generalization studies are conducted. *British Journal of Mathematical and Statistical Psychology*, n/a-n/a. https://doi.org/10.1111/j.2044-8317.2012.02057.x

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

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

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

Sandoval-Lentisco, A., López-Nicolás, R., López-López, J. A., & Sánchez-Meca, J. (2023). Florida Obsessive-Compulsive Inventory and Children's Florida Obsessive Compulsive Inventory: A reliability generalization meta-analysis. *Journal of Clinical Psychology*, *79*(1), 28–42. https://doi.org/10.1002/jclp.23416

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

Schmid, C. H., Stijnen, T., & White, I. R. (Eds.). (2020). *Handbook of meta-analysis* (First edition). Taylor and Francis.

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

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

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

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Spearman, C. (1907). Demonstration of Formulæ for True Measurement of Correlation. *American Journal of Psychology*, *18*(2), 161–169.

Spearman, C. (1913). Correlations of sums or differences. *British Journal of Psychology*, *5*(4), 417–426. https://doi.org/10.1111/j.2044-8295.1913.tb00072.x

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

Storch, E. A., Bagner, D., Merlo, L. J., Shapira, N. A., Geffken, G. R., Murphy, T. K., & Goodman, W. K. (2007). Florida obsessive-compulsive inventory: Development, reliability, and validity. *Journal of Clinical Psychology*, *63*(9), 851–859. https://doi.org/10.1002/jclp.20382

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS: Robust variance estimation. *Research Synthesis Methods*, *5*(1), 13–30. https://doi.org/10.1002/jrsm.1091

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

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is Datametrics: The Test is not Reliable. *Educational and Psychological Measurement*, *60*(2), 174–195. https://doi.org/10.1177/0013164400602002

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

Vacha-Haase, T. (1998). Reliability Generalization: Exploring Variance in Measurement Error Affecting Score Reliability Across Studies. *Educational and Psychological Measurement*, *58*(1), 6–20. https://doi.org/10.1177/0013164498058001002

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods, 45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods, 47*(4), 1274–1294. https://doi.org/10.3758/s13428-014-0527-2

Van Oppen, P., Hoekstra, R. J., & Emmelkamp, P. M. G. (1995). The structure of obsessive-compulsive symptoms. *Behaviour Research and Therapy, 33*(1), 15–23. https://doi.org/10.1016/0005-7967(94)E0010-G

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

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

Vicent, M., Rubio-Aparicio, M., Sánchez-Meca, J., & Gonzálvez, C. (2019). A reliability generalization meta-analysis of the child and adolescent perfectionism scale. *Journal of Affective Disorders, 245,* 533–544. https://doi.org/10.1016/j.jad.2018.11.049

Viechtbauer, W. (2005). Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. *Journal of Educational and Behavioral Statistics, 30*(3), 261–293. https://doi.org/10.3102/10769986030003261

Viechtbauer, W. (2010). Conducting Meta-Analyses in *R* with the **metafor** Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

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

Wickham, H., & RStudio. (2023). *forcats: Tools for Working with Categorical Variables (Factors)* (1.0.0) [R]. https://forcats.tidyverse.org/

Wilke, C. O. (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'* (1.1.1) [R]. https://wilkelab.org/cowplot/

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

Yin, P., & Fan, X. (2000). Assessing the Reliability of Beck Depression Inventory Scores: Reliability Generalization Across Studies. *Educational and Psychological Measurement*, *60*(2), 201–223.

Appendices

Appendix 2A:

Supplementary Tables and Figures for Chapter 2

Tables

Table 2A.1

Mathematical formulation of the three statistical models proposed in RG meta-analysis.

	Mathematical model	**Parameter to estimate**
Fixed-effect model	$\hat{\theta}_i = \theta + e_i$	θ
Varying-coefficient model	$\hat{\theta}_i = \theta_i + e_i$	$k^{-1}\sum_{i=1}^{k}\theta_i$
Random-effects model	$\hat{\theta}_i = \mu_\theta + e_i + \varepsilon_i$	μ_θ

Note: $\hat{\theta}_i$: reliability estimate reported in the i_{th} study. θ: population reliability coefficient common to all individual reliability estimates when assuming a FE model. e_i: sampling error of the i_{th} reliability estimate. μ_θ: pooled parametric reliability coefficient when assuming an RE model. ε_i: error due to sampling of population reliability coefficients.

Table 2A.2

Statistical methods to calculate an average reliability coefficient.

	Statistical Model				
Transformation method	**OLS**	**FE**	**VC**	**RE/Rei**	**REn**
No transformation	✓	✓	✓[a]	✓	✓
Fisher’s Z	✓	✓	-	✓	-
Hakstian and Whalen	✓	✓	-	✓	-
Bonett	✓	✓	-	✓	-

Note: OLS: Ordinary Least Squares (unweighted conventional statistical methods). FE: Fixed-effect model. VC: Varying-Coefficient model. RE: Standard Random-Effects model weighting by the inverse variance. REi: Random-Effects model with the improved method of Hartung and Knapp (2001). REn: Random-effects model weighting by sample size. [a]Note that the average reliability coefficient calculated under the VC model coincides with that of the OLS model for untransformed reliability coefficients.

Table 2A.3

Methods to construct a confidence interval around the average reliability coefficient.

	Statistical Model					
Transformation method	**OLS**	**FE**	**VC**	**RE**	**REi**	**REn**
No transformation	✓	✓	-	✓	✓	✓
Fisher’s Z	✓	✓	-	✓	✓	-
Hakstian and Whalen	✓	✓	-	✓	✓	-
Bonett	✓	✓	✓	✓	✓	-

Note: OLS = Ordinary Least Squares (unweighted conventional statistical methods). FE = Fixed-Effect model. VC = Varying-Coefficient model. RE = standard Random-Effects model weighting by the inverse variance. REi = Random-Effects model with the improved method of Hartung and Knapp (2001). REn: Random-Effects model weighting by sample size.

Table 2A.4

Descriptive statistics of number of studies included in each meta-analysis, sample sizes, kurtosis and skewness.

		Mean	SD	Min	Q1	Median	Q3	Max	Range
	Number of Studies	31.34	47.37	5	9	14	40	319	314
	Sample Size	209.19	107.03	38	125.25	220	249.38	799	761
Kurtosis	**No Transformation**	3.74	2.53	1.42	2.21	2.95	4.15	18.48	17.06
	Fisher's Z	2.98	1.21	1.52	2.18	2.61	3.44	8.32	6.81
	Hakstian-Whalen's	3.01	1.36	1.49	2.16	2.61	3.26	9.44	7.95
	Bonett's	2.94	1.14	1.53	2.15	2.59	3.34	7.00	5.47
Skewness	**No Transformation**	-.75	.85	-3.47	-1.18	-.71	-.2	1.21	4.69
	Fisher's Z	.01	.73	-1.95	-.42	.07	.48	2.29	4.24
	Hakstian-Whalen's	.2	.73	-1.6	-.29	.14	.63	2.22	3.81
	Bonett's	-.09	.72	-2.3	-.55	-.12	.33	1.84	4.14

Note: SD: Standard Deviation. Min. and Max.: Minimum and Maximum. Q1 and Q3: quartiles 1 and 3.

Table 2A.5

Post-hoc comparisons between different transformations of the coefficients regarding to skewness

Transformation	Transformation	Mean Difference	*SE*	$p_{Bonferroni}$
No Transformation	**Fisher's Z**	-.756	.040	< .001
	Hakstian-Whalen	-.948	.132	< .001
	Bonett	-.664	.126	< .001
Fisher's Z	**Hakstian-Whalen**	-.192	.124	.74
	Bonett	.092	.123	1
Hakstian-Whalen	**Bonett**	.284	.014	< .001

Note. SE = Standard Error of the mean difference.

Table 2A.6

Post-hoc comparisons between different transformations of the coefficients regarding to kurtosis

Transformation	Transformation	Mean Difference	*SE*	$p_{Bonferroni}$
No Transformation	**Fisher's Z**	.769	.144	< .001
	Hakstian-Whalen	.736	.116	< .001
	Bonett	.802	.160	< .001
Fisher's Z	**Hakstian-Whalen**	-.033	.034	1
	Bonett	.033	.020	.548
Hakstian-Whalen	**Bonett**	.066	.050	1

Note. *SE* = Standard Error of the mean difference.

Table 2A.7

Descriptive statistics of average alpha coefficients for each τ^2 estimator

τ^2	Transformation	Mean	SD	Min	Q1	Median	Q3	Max	Range
DL	**No Transformation**	.831	.07	.622	.791	.837	.883	.975	.353
	Fisher's Z	.833	.069	.62	.787	.84	.887	.986	.366
	Hakstian-Whalen	.832	.069	.622	.789	.838	.885	.98	.358
	Bonett	.834	.068	.624	.788	.842	.887	.986	.361
REML	**No Transformation**	.828	.069	.616	.791	.832	.877	.975	.359
	Fisher's Z	.833	.069	.619	.786	.841	.887	.986	.366
	Hakstian-Whalen	.831	.069	.62	.789	.838	.884	.98	.36
	Bonett	.834	.068	.624	.789	.842	.887	.986	.361

Note. DL= DerSimonian-Laird estimator. REML= Restricted Maximum-Likelihood estimator. *SD*: Standard Deviation. Min. and Max.: Minimum and Maximum. *Q1* and *Q3*: quartiles 1 and 3.

Table 2A.8

Results of the repeated measures ANOVA to calculate the average alpha coefficient

	Sum of Squares	***df***	**Mean Square**	***F***	***p***	**η^2**
Transformation	.004	3	.001	43.43	< .001	.241
Residual	.011	411	.000			
τ^2 estimator	.000	1	.000	1.11	.294	.008
Residual	.022	137	.000			
Transformation * τ^2 estimator	.000	3	.000	26.29	< .001	.161
Residual	.002	411	.000			

Table 2A.9

Post-hoc comparisons for the interaction between □^{2} estimator and transformation of the coefficients regarding the average alpha coefficient.

τ^2	**Transformation**	**τ^2**	**Transformation**	**Mean Difference**	***SE***	**$p_{\text{Bonferroni}}$**
DL	**No Transformation**	**REML**	**No Transformation**	.003	.001	.037
	Fisher's Z		**Fisher's Z**	.000	.001	1
	Hakstian-Whalen		**Hakstian-Whalen**	.000	.001	1
	Bonett		**Bonett**	.000	.001	1

Note. DL= DerSimonian-Laird estimator. REML= Restricted Maximum-Likelihood estimator. Only those combinations that were of interest for the study have been included in the table.

Table 2A.10

Descriptive statistics of confidence width for each random-effects model and τ^2 estimator.

τ^2	Model	Transformation	Mean	SD	Min	Q_1	Median	Q_3	Max	Range
DL	RE	No Transformation	.059	.052	.009	.03	.043	.072	.412	.403
		Fisher's Z	.07	.057	.01	.035	.054	.089	.421	.411
		Hakstian-Whalen	.068	.059	.01	.034	.051	.086	.46	.45
		Bonett	.069	.058	.01	.034	.052	.089	.417	.407
	REi	No Transformation	.079	.071	.011	.035	.059	.099	.543	.532
		Fisher's Z	.084	.077	.012	.037	.058	.107	.589	.578
		Hakstian-Whalen	.082	.075	.012	.036	.06	.104	.573	.561
		Bonett	.084	.08	.012	.037	.058	.107	.637	.625
REML	RE	No Transformation	.069	.055	.011	.032	.054	.083	.387	.376
		Fisher's Z	.071	.056	.011	.037	.054	.089	.412	.401
		Hakstian-Whalen	.07	.055	.011	.035	.054	.087	.402	.39
		Bonett	.071	.057	.012	.037	.052	.088	.422	.411
	REi	No Transformation	.082	.075	.012	.035	.061	.1	.544	.532
		Fisher's Z	.084	.077	.012	.038	.058	.11	.589	.577
		Hakstian-Whalen	.083	.075	.012	.037	.06	.107	.574	.562
		Bonett	.084	.08	.012	.038	.058	.109	.637	.625

Note. DL= DerSimonian-Laird estimator. REML= Restricted Maximum-Likelihood estimator. RE= Random-Effects model. REi: Random-Effects model with the improved method of Hartung and Knapp (2001). *SD*: Standard Deviation. Min. and Max.: Minimum and Maximum. Q_1 and Q_3: quartiles 1 and 3.

Table 2A.11

Results of the repeated measures ANOVA to calculate the confidence width.

	Sum of Squares	***df***	**Mean Square**	***F***	***p***	***η*²**
Transformation	.007	3	.002	5.81	< .001	.041
Residual	.17	411	0			
τ^2	.032	1	.032	2.12	.147	.015
Residual	2.072	137	.015			
Model	.105	1	.105	41.39	< .001	.232
Residual	.348	137	.003			
Transformation * τ^2	0	3	.002	4.5	.004	.032
Residual	.16	411	0			
Transformation * Model	.002	3	0	2.19	.088	.016
Residual	.15	411	0			
τ^2 * Model	.003	1	.003	5.72	.018	.04
Residual	.062	137	0			
Transformation * τ^2 * Model	.001	3	0	1.18	.316	.009
Residual	.134	411	0			

Table 2A.12

Post-hoc comparisons for the interaction between τ² *estimator and the transformation of the coefficients regarding the confidence width.*

τ^2	**Transformation**	τ^2	**Transformation**	**Mean Difference**	***SE***	$p_{\text{Bonferroni}}$
DL	**No Transformation**	**REML**	**No Transformation**	-.013	.006	1
	Fisher's Z		**Fisher's Z**	-.007	.006	1
	Hakstian-Whalen		**Hakstian-Whalen**	-.007	.006	1
	Bonett		**Bonett**	-.004	.003	1

Note. DL= DerSimonian-Laird estimator. REML= Restricted Maximum-Likelihood estimator. Only those combinations that were of interest for the study have been included in the table.

Table 2A.13

Post-hoc comparisons for the interaction between τ^2 estimator and the two random effects models regarding the confidence width.

τ^2	**Model**	**τ^2**	**Model**	**Mean Difference**	***SE***	**p$_{\text{Bonferroni}}$**
DL	RE	REML	RE	-.01	.006	.643
	REi		REi	-.005	.005	1

Note. DL= DerSimonian-Laird estimator. REML= Restrictied máximum-likelihood estimator. RE= Standard Random-Effects model weighting by the inverse variance. REi= Random-Effects model with the improved method of Hartung and Knapp (2001). Only those combinations that were of interest for the study have been included in the table.

Table 2A.14

Post-hoc comparisons between statistical models to calculate average alpha

Model	**Model**	**Mean Difference**	***SE***	***P$_{Bonferroni}$***
FE	**OLS**	.019	.004	< .001
	RE	.015	.004	.003
	REn	.021	.007	.008
OLS	**RE**	-.004	.004	1
	REn	.002	.007	1
RE	**REn**	.007	.007	1

Note. FE: Fixed-Effect model. OLS: Ordinary Least Squares. RE: Random-Effects model. REn: Random-Effects model weighted by sample size. *SE*: Standard Error.

Table 2A.15

Bonferroni's post hoc comparisons for analytic strategy to calculate average alpha coefficient.

Model	Transformation	Model	Transformation	Mean Difference	*SE*	*t*	$p_{\text{Bonf.}}$
FE	**Bonett**	**FE**	**Hakstian-Whalen**	-.011	.008	-1.277	1
			No Transformation	-.029	.008	-3.48	.04
			Fisher's Z	.001	.008	.148	1
		OLS	**Bonett**	.004	.008	.48	1
		RE	**Bonett**	.003	.008	.393	1
FE	**Hakstian-Whalen**	**FE**	**No Transformation**	-.019	.008	-2.203	1
			Fisher's Z	.012	.008	1.425	1
		OLS	**Bonett**	.015	.008	1.757	1
			Hakstian-Whalen'	.02	.008	2.32	1
		RE	**Hakstian-Whalen**	.016	.008	1.94	1
FE	**No Transformation**	**FE**	**Fisher's Z**	.031	.008	3.628	.023
		OLS	**No Transformation**	.048	.008	5.67	< .001
		RE	**No Transformation**	.036	.008	4.275	.002
		REn	**No Transformation**	.041	.008	4.86	< .001
FE	**Fisher's Z**	**OLS**	**Fisher's Z**	.004	.008	.509	1
		RE	**Fisher's Z**	.003	.008	.395	1
OLS	**Bonett**	**OLS**	**Hakstian-Whalen**	.005	.008	.563	1
			No Transformation	.014	.008	1.71	1
			Fisher's Z	.001	.008	.177	1
		RE	**Bonett**	-.000	.008	-.087	1
OLS	**Hakstian-Whalen**	**OLS**	**No Transformation**	.01	.008	1.147	1
			Fisher's Z	-.003	.008	-.386	1
		RE	**Hakstian-Whalen**	-.003	.008	-.38	1
OLS	**No Transformation**	**OLS**	**Fisher's Z**	-.013	.008	-1.533	1
		RE	**No Transformation**	-.012	.008	-1.395	1
		REn	**No Transformation**	-.007	.008	-.811	1
OLS	**Fisher's Z**	**RE**	**Fisher's Z**	-.000	.008	-.114	1
RE	**No Transformation**	**REn**	**No Transformation**	.005	.008	.585	1

Note. FE: Fixed-Effect Model. OLS: Ordinary Least Squares. RE: Random Effects Model. REn: Random Effects Model weighted by sample size. NT: untransformed reliability coefficients. Z: Fisher's Z transformation. HW: Hakstian and Whalen's transformation. B: Bonett's transformation. *SE*: Standard Error. Only those combinations that were of interest for the study have been included in the table.

Table 2A.16

Post-hoc comparisons between statistical models to calculate the confidence width of average alpha coefficient

Model	**Model**	**Mean Difference**	***SE***	$p_{\text{bonferroni}}$
FE	**OLS**	-.070	.004	< .001
	RE	-.051	.004	< .001
	REi	-.067	.004	< .001
	REn	-.047	.006	< .001
	VC	-.009	.006	1
OLS	**RE**	.02	.004	< .001
	REi	.004	.004	1
	REn	.024	.006	< .001
	VC	.061	.006	< .001
RE	**REi**	-.016	.004	< .001
	REn	.004	.006	1
	VC	.041	.006	< .001
REi	**REn**	.197	.006	.007
	VC	.057	.006	< .001
REn	**VC**	.038	.007	< .001

Note. FE: Fixed-Effect Model. OLS: Ordinary Least Squares. RE: Random Effects Model. REi: Random-Effects model with the improved method of Hartung and Knapp (2001). REn: Random-Effects model weighting by sample size. VC: Varying-Coefficient model. SE: Standard Error.

Table 2A.17

Post-hoc comparisons between the transformation of the coefficients regarding the I^2 index

Transformation	Transformation	Mean Difference	*SE*	$p_{\text{Bonferroni}}$
No Transformation	**Fisher's Z**	2.62	.396	< .001
	Hakstian-Whalen	-.868	.163	< .001
	Bonett	-.86	.22	< .001
Fisher's Z	**Hakstian-Whalen**	-3.488	.361	< .001
	Bonett	-3.481	.351	< .001
Hakstian-Whalen	**Bonett**	.008	.07	1

Table 2A.18

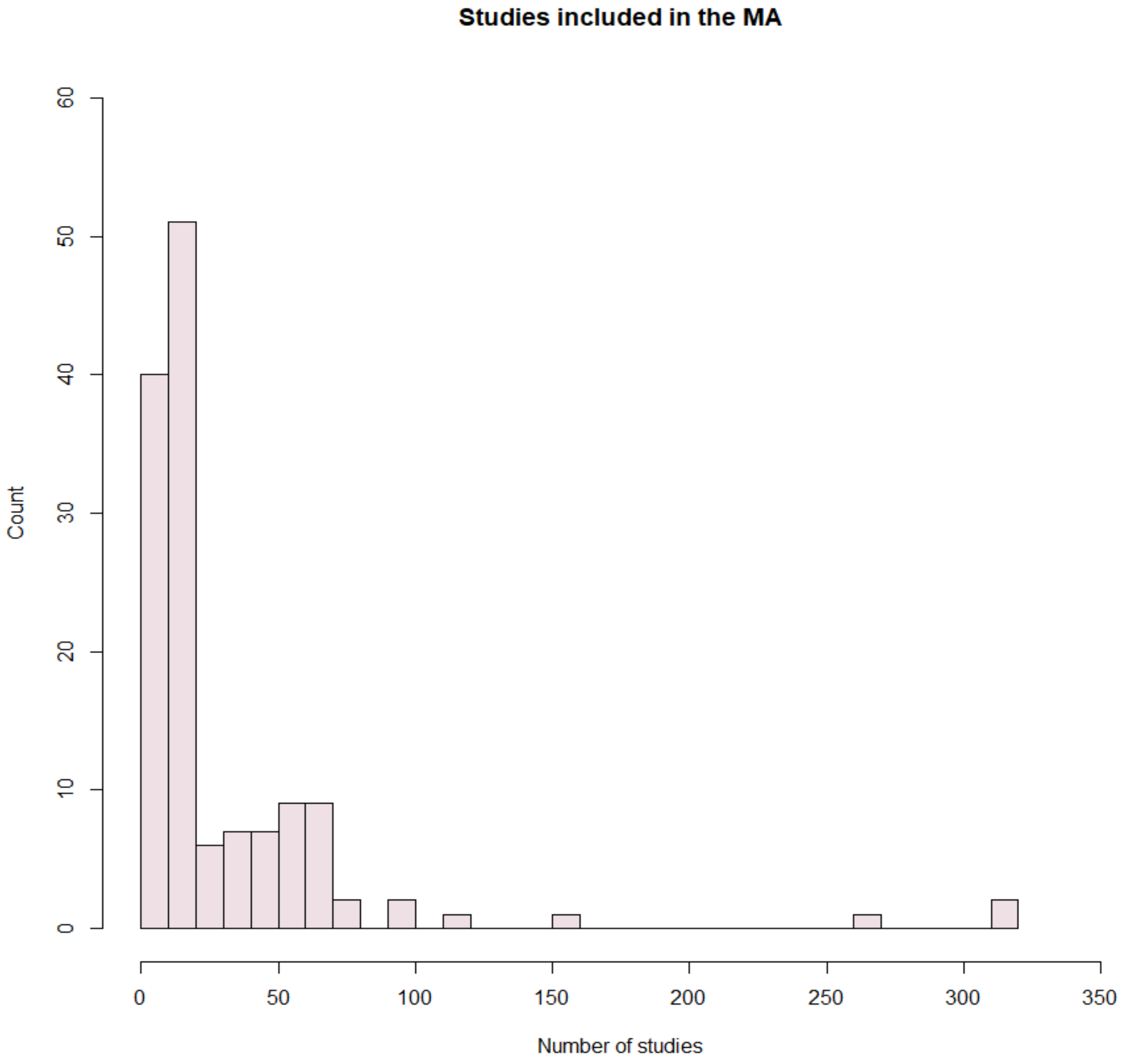
Post-hoc comparisons between transformation of the coefficients regarding prediction intervals width.

Transformation	Transformation	Mean Difference	*SE*	$p_{\text{Bonferroni}}$
No Transformation	**Fisher's Z**	-.066	.009	< .001
	Hakstian-Whalen	-.047	.005	< .001
	Bonett	-.091	.013	< .001
Fisher's Z	**Hakstian-Whalen**	.019	.001	.009
	Bonett	-.024	.005	< .001
Hakstian-Whalen	**Bonett**	-.043	.009	< .001

Figures

Figure 2A.1

Distribution of the number of studies for the 138 RG datasets.



Appendix 2B:

Full search strategies and screening process summary

Table 2B.1

Full search strategy for each database

Database	Search strategy
SCOPUS	(TITLE-ABS-KEY (“reliability generalization”) OR TITLE-ABS-KEY (“meta analysis of internal consistence”) OR TITLE-ABS-KEY ("meta analysis of alpha coefficients")) AND PUBYEAR > 1997 AND PUBYEAR < 2021
Google Scholar	allintitle: "reliability generalization" "meta analysis of internal consistency" "meta analysis of alpha coefficients" Range: 1998-2020
EBSCOHOST (MEDLINE, APA PscycInfo, Education Sourse, APA PsycArticles, Gender Studies Database, PSICODOC)	TI “Reliability Generalization Meta-Analysis” OR AB “Reliability Generalization Meta-Analysis” OR TI “Meta-Analysis of Internal Consistence” OR AB “Meta-Analysis of Internal Consistence” OR TI “Meta-Analysis of Alpha Coefficients” OR AB “Meta-Analysis of Alpha Coefficients”

Appendix 2C:

References of the Reliability Generalization Meta-analyses included in the Meta-Review

Aguayo, R., Vargas, C., de la Fuente, E. I., & Lozano, L. M. (2011). A meta-analytic reliability generalization study of the Maslach Burnout Inventory. *International Journal of Clinical and Health Psychology*, *11*(2), 343-361.

Barlow, K. M., & Zangaro, G. A. (2010). Meta-analysis of the reliability and validity of the Anticipated Turnover Scale across studies of registered nurses in the United States. *Journal of Nursing Management*, *18*, 862-873. https://doi.org/10.1111/j.1365-2834.2010.01171.x

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

Deng, J., Wang, M.-C., Zhang, X., Shou, Y., Gao, Y., & Luo, J. (2019). The inventory of callous unemotional traits: A reliability generalization meta-analysis. *Psychological Assessment*, *31*(6), 765-780. https://doi.org/10.1037/pas0000698

Hallinger, P., Wang, W.-C., & Chen, C.-W. (s. f.). Instructional management rating scale: A meta-analysis of reliability studies. *Educational Administration Quarterly*, *49*(2), 272-309. https://doi.org/10.1177/0013161X12468149

Hart, P. D., & Kang, M. (s. f.). Reliability of the short-form health survey (SF-36) in physical activity research using meta-analysis. *World Journal of Preventive Medicine*, *3*(2), 17-23. https://doi.org/10.12691/jpm-3-2-1

Huynh, Q.-L., Howell, R. T., & Benet-Martínez, V. (2009). Reliability of bidimensional acculturation scores. *Journal of Cross-Cultural Psychology*, *40*(2), 256-274. https://doi.org/10.1177/0022022108328919

Khaleque, A., & Rohner, R. P. (2002). Reliability of measures assessing the pancultural association between perceived parental acceptance-rejection and psychological adjustment: A meta-analysis of cross-cultural and intracultural studies. *Journal of Cross-Cultural Psychology*, *33*(1), 87-99. https://doi.org/10.1177/0022022102033001006

Lee, C.-P., Chiu, Y.-W., Chu, C.-L., Chen, Y., Jiang, K.-H., Chen, J.-L., & Chen, C.-Y. (2016). A reliability generalization meta-analysis of coefficient alpha and test–retest coefficient for the aging males' symptoms (AMS) scale. *The Aging Male*, *19*(4), 244-253. https://doi.org/10.1080/13685538.2016.1246525

López-Nicolás, R., Rubio-Aparicio, M., López-Ibáñez, C. y Sánchez-Meca, J. (in press). A reliability generalization meta-analysis of the Dimensional Obsessive-Compulsive Scale. *Psicothema*.

López-Pina, José A, Sánchez-Meca, J., & Rosa-Alcázar, A. I. (2009). The Hamilton rating scale for depression: A meta-analytic reliability generalization study. *International Journal of Clinical and Health Psychology*, *9*(1), 143-159. https://doi.org/10.7334/psicothema2020.455

López-Pina, José Antonio, Sánchez-Meca, J., López-López, J. A., Marín-Martínez, F., Núñez-Núñez, R. M., Rosa-Alcázar, A. I., Gómez-Conesa, A., & Ferrer-Requena,

J. (2015a). The Yale–Brown obsessive compulsive scale: A reliability generalization meta-analysis. *Assessment*, *22*(5), 619-628. https://doi.org/10.1177/1073191114551954

López-Pina, José Antonio, Sánchez-Meca, J., López-López, J. A., Marín-Martínez, F., Núñez-Núñez, R. M., Rosa-Alcázar, A. I., Gómez-Conesa, A., & Ferrer-Requena, J. (2015b). Reliability Generalization Study of the Yale–Brown Obsessive–Compulsive Scale for Children and Adolescents. *Journal of Personality Assessment*, *97*(1), 42-54. https://doi.org/10.1080/00223891.2014.930470

Meseguer-Henarejos, A.-B., Rubio-Aparicio, M., López-Pina, J.-A., Carles-Hernández, R., & Gómez-Conesa, A. (2019). Characteristics that affect score reliability in the Berg Balance Scale: A meta-analytic reliability generalization study. *European Journal of Physical and Rehabilitation Medicine*, *55*(5), 570-584. https://doi.org/10.23736/S1973-9087.19.05363-2

Núñez-Núñez, R. M., Rubio-Aparicio, M., Marín-Martínez, F., Sánchez-Meca, J., López-Pina, J. A., & López-López, J. A. (2022). A Reliability Generalization Meta-Analysis of the Padua Inventory-Revised (PI-R). *International Journal of Clinical and Health Psychology*, *22*(1), 1-12. https://doi.org/10.1016/j.ijchp.2021.100277

Phillips, C. E., King, C., Kivisalu, T. M., & O'Toole, S. K. (2016). A reliability generalization of the Suinn-Lew asian self-identity acculturation scale. *SAGE Open*, 1-15. https://doi.org/10.1177/2158244016661748

Piqueras, J. A., Martín-Vivar, M., Sandin, B., San Luis, C., & Pineda, D. (2017). The Revised Child Anxiety and Depression Scale: A systematic review and reliability generalization meta-analysis. *Journal of Affective Disorders*, *218*, 153-169. https://doi.org/10.1016/j.jad.2017.04.022

Rohner, R. P., & Khaleque, A. (2003). Reliability and Validity of the Parental Control Scale: A Meta-Analysis of Cross-Cultural and Intracultural Studies. *Journal of Cross-Cultural Psychology*, *34*(6), 643-649. https://doi.org/10.1177/0022022103255650

Rouse, S. V. (2007). Using reliability generalization methods to explore measurement error: An illustration using the MMPI–2 PSY–5 scales. *Journal of Personality Assessment*, *88*(3), 264-275. https://doi.org/10.1080/00223890701293908

Rubio-Aparicio, Maria, Badenes-Ribera, L., Sánchez-Meca, J., Fabris, M. A., & Longobardi, C. (2020). A reliability generalization meta-analysis of self-report measures of muscle dysmorphias. *Clinical Psychology: Science and Practice*, 1-24. https://doi.org/10.1111/cpsp.12303

Rubio-Aparicio, María, Núñez-Núñez, R. M., Sánchez-Meca, J., López-Pina, J. A., Marín-Martínez, F., & López-López, J. A. (2020). The Padua inventory–Washington State University revision of obsessions and compulsions: A reliability generalization meta-analysis. *Journal of Personality Assessment*, *102*(1), 113-123. https://doi.org/10.1080/00223891.2018.1483378

Sánchez-Meca, J., Alacid-de-Pascual, I., López-Pina, J. A., & Sánchez-Jiménez, J. de la C. (2016). Meta-análisis de generalización de la fiabilidad del inventario de obsesiones de Leyton versión para niños auto-aplicada. *Revista Española de Salud Pública*, *90*, 1-14.

Sánchez-Meca, J., López-Pina, J. A., López-López, J. A., Marín-Martínez, F., Rosa-Alcázar, A. I., & Gómez-Conesa, A. (2011). The Maudsley obsessive-compulsive inventory: A reliability generalization meta-analysis. *International Journal of Clinical and Health Psychology*, *11*(3), 473-493.

Sánchez-Meca, J., Rubio-Aparicio, M., Núñez-Núñez, R. M., López-Pina, J., Marín-Martínez, F., & López-López, J. A. (2017). A reliability generalization meta-analysis of the padua inventory of obsessions and compulsions. *The Spanish Journal of Psychology*, *20*(70), 1-15. https://doi.org/10.1017/sjp.2017.65

Sandoval-Lentisco, A., López-Nicolás, R., López-López, JA., & Sánchez-Meca, J. (2022). Florida Obsessive-Compulsive Inventory and Children's Florida Obsessive Compulsive Inventory: A reliability generalization meta-analysis. *Journal of Clinical Psychology*, *79*, 28-42. https://doi.org/10.1002/jclp.23416

Scaini, S., Battaglia, M., Beidel, D. C., & Ogliari, A. (2012). A meta-analysis of the cross-cultural psychometric properties of the Social Phobia and Anxiety Inventory for Children (SPAI-C). *Journal of Anxiety Disorders*, *26*, 182-188. https://doi.org/10.1016/j.janxdis.2011.11.002

Stockings, E., Degenhardt, L., Lee, Y. Y., Mihalopoulos, C., Liu, A., Hobbs, M., & Patton, G. (2015). Symptom screening scales for detecting major depressive disorder in children and adolescents: A systematic review and meta-analysis of reliability, validity and diagnostic utility. *Journal of Affective Disorders*, *174*, 447-463. https://doi.org/10.1016/j.jad.2014.11.061

University of Massachusetts, Hess, T. J., McNab, A. L., Niagara University, Basoglu, K. A., & University of Delaware. (2014). Reliability Generalization of Perceived Ease of Use, Perceived Usefulness, and Behavioral Intentions. *MIS Quarterly*, *38*(1), 1-28. https://doi.org/10.25300/MISQ/2014/38.1.01

Vicent, M., Rubio-Aparicio, M., Sánchez-Meca, J., & Gonzálvez, C. (2019). A reliability generalization meta-analysis of the child and adolescent perfectionism scale.

Journal of Affective Disorders, *245*, 533-544. https://doi.org/10.1016/j.jad.2018.11.049

Vilagut, G., Ferrer, M., Rajmil, L., Rebollo, P., Permanyer-Miralda, G., Quintana, J. M., Santed, R., Valderas, J. M., Ribera, A., Domingo-Salvany, A., & Alonso, J. (2005). El cuestionario de salud SF-36 español: Una década de experiencia y nuevos desarrollos. *Gaceta Sanitaria*, *19*(2), 135-150. https://doi.org/10.1157/13074369

Warne, R. T. (2011). A reliability generalization of the overexcitability questionnaire–two. *Journal of Advanced Academics*, *22*(5), 671-692. https://doi.org/10.1177/1932202X11424881

Zangaro, G. A., & Soeken, K. L. (2005). Meta-analysis of the reliability and validity of part b of the index of work satisfaction across studies. *Journal of Nursing Measurement*, *13*(1), 1-16.

Appendix 4A:

Supplementary Tables for Chapter 4

Average alpha coefficients

Table 4A.1

Summary table of the comparisons established with respect to the calculation of average alpha coefficients.

Model (I)	Transformation (I)	Model (II)	Transformation (II)
Conventional	Not Transformed	**Multilevel**	Not Transformed
Conventional	Bonett's	**Multilevel**	Bonett's
Conventional	Not Transformed	**Homoscedastic**	Not Transformed
Conventional	Not Transformed	**Heteroscedastic**	Not Transformed
Conventional	Bonett's	**Homoscedastic**	Bonett's
Conventional	Bonett's	**Heteroscedastic**	Bonett's
Homoscedastic	Not Transformed	**Heteroscedastic**	Not Transformed
Homoscedastic	Bonett's	**Heteroscedastic**	Bonett's
Conventional	Not Transformed	**Conventional**	Bonett's
Multilevel	Not Transformed	**Multilevel**	Bonett's

Note: the reference value is the combination of the first and second columns

Confidence width

Table 4A.2

Summary table of the first comparison established with respect to the calculation of confidence width.

Model (I)	Transformation (I)	Confidence Width method (I)	Model (II)	Transformation (II)	Confidence Width method (II)
Conventional	Not Transformed	Standard	**Multilevel**	Not Transformed	Standard
Conventional	Bonett's	Standard	**Multilevel**	Bonett's	Standard
Conventional	Not Transformed	Knapp-Hartung	**Multilevel**	Not Transformed	Knapp-Hartung
Conventional	Bonett's	Knapp-Hartung	**Multilevel**	Bonett's	Knapp-Hartung
Conventional	Not Transformed	-	**RVE**	Not Transformed	-
Conventional	Bonett's	-	**RVE**	Bonett's	-

Note: the reference value is the combination of the first three columns

Table 4A.3

Summary table of the second comparison established with respect to the calculation of confidence width.

Model (I)	Transformation (I)	Confidence Width method (I)	Model (II)	Transformation (II)	Confidence Width method (II)
Conventional	Not Transformed	Standard	**Homoscedastic**	Not Transformed	Not Transformed
Conventional	Not Transformed	Standard	**Heteroscedastic**	Not Transformed	Not Transformed
Conventional	Bonett's	Standard	**Homoscedastic**	Bonett's	Standard

Conventional	Bonett’s	Standard	**Heteroscedastic**	Bonett’s	Standard
Conventional	Not Transformed	Knapp-Hartung	**Homoscedastic**	Not Transformed	Knapp-Hartung
Conventional	Not Transformed	Knapp-Hartung	**Heteroscedastic**	Not Transformed	Knapp-Hartung
Conventional	Bonett’s	Knapp-Hartung	**Homoscedastic**	Bonett’s	Knapp-Hartung
Conventional	Bonett’s	Knapp-Hartung	**Heteroscedastic**	Bonett’s	Knapp-Hartung

Note: the reference value is the combination of the first three columns

Table 4A.4

Summary table of the third comparison established with respect to the calculation of confidence width.

Model (I)	Transformation (I)	Confidence Width method (I)	Model (II)	Transformation (II)	Confidence Width method (II)
Homoscedastic	Not Transformed	Standard	**Heteroscedastic**	Not Transformed	Standard
Homoscedastic	Bonett’s	Standard	**Heteroscedastic**	Bonett’s	Standard
Homoscedastic	Not Transformed	Knapp-Hartung	**Heteroscedastic**	Not Transformed	Knapp-Hartung
Homoscedastic	Bonett’s	Knapp-Hartung	**Heteroscedastic**	Bonett’s	Knapp-Hartung

Note: the reference value is the combination of the first three columns

Table 4A.5

Summary table of the fourth comparison established with respect to the calculation of confidence width.

Model (I)	Transformation (I)	Confidence Width method (I)	Model (II)	Transformation (II)	Confidence Width method (II)
Conventional	Not Transformed	Standard	**Conventional**	Not Transformed	Knapp-Hartung
Conventional	Bonett’s	Standard	**Conventional**	Bonett’s	Knapp-Hartung
Multilevel	Not Transformed	Standard	**Multilevel**	Not Transformed	Knapp-Hartung
Multilevel	Bonett’s	Standard	**Multilevel**	Bonett’s	Knapp-Hartung

Note: the reference value is the combination of the first three columns

Table 4A.6

Summary table of the fifth comparison established with respect to the calculation of confidence width.

Model (I)	Transformation (I)	Confidence Width method (I)	Model (II)	Transformation (II)	Confidence Width method (II)
Conventional	Not Transformed	Standard	**Conventional**	Bonett’s	Standard
Conventional	Not Transformed	Knapp-Hartung	**Conventional**	Bonett’s	Knapp-Hartung
Multilevel	Not Transformed	Standard	**Multilevel**	Bonett’s	Standard
Multilevel	Not Transformed	Knapp-Hartung	**Multilevel**	Bonett’s	Knapp-Hartung

Note: the reference value is the combination of the first three columns