informaciópsicològica
I.S.S.N. publicación impresa: 0214-347 X
doi: dx.medra.org/10.14635/IPSIC.2020.119.6

dossier
Novedades en instrumentos de evaluación psicológica

I.S.S.N.: 1989-9076
(2020) 119, 17-32

Meta-análisis de generalización de la fiabilidad

Reliability Generalization Meta-Analysis

Fecha de recepción: 21-01-2020
Fecha de aceptación: 18-06-2020

Laura Badenes-Ribera
Universitat de València
María Rubio-Aparicio
Universitat d'Alacant
Julio Sánchez-Meca
Universidad de Murcia

resumen/abstract:

Un meta-análisis de generalización de la fiabilidad (MA GF) es un método para integrar estadísticamente las estimaciones de fiabilidad obtenidas en diferentes aplicaciones de un test. El MA GF permite a los investigadores caracterizar la fiabilidad promedio de las puntuaciones obtenidas en un test en múltiples estudios y situaciones y estimar el grado de variabilidad en los coeficientes de fiabilidad en diferentes tipos de medidas, muestras y contextos. Por lo tanto, sus resultados permiten ofrecer pautas a los investigadores y profesionales aplicados sobre qué escalas son más fiables para evaluar un constructo y en qué circunstancias. Así pues, los investigadores y profesionales necesitan saber qué son los MA GF, cómo se hacen y, lo que es más importante, cómo podemos hacer valoraciones críticas de ellos. El propósito de este artículo es presentar los MA GF, así como una guía orientativa sobre cómo hacer una lectura crítica de ellos. Para ello, un reciente MA GF es utilizado para ilustrar la guía propuesta. Finalmente, se presentan algunas observaciones finales.

A reliability generalization meta-analysis (RG MA) is a suitable method to statistically integrate the reliability estimates obtained in different applications of a test. RG MA allows researchers to characterize the average reliability of scores obtained by a test across multiple studies and situations and estimate the degree of variability in reliability coefficients across different types of measures, samples, and contexts. Therefore, its results enable us to offer guidelines to applied researchers and practitioners about which scales are more reliable for assessing a constructe and in what circumstances. Thus, applied researchers and practitioners need to know what RG MAs are, how they are done and, most importantly, how we can carry out critical appraisal of RG MAs. The purpose of this article is to present RG MAs, together with some guidelines to warrant a critical reading of them. For this, a recent reliability generalization meta-analysis is used to illustrate the guidelines proposed. Finally, some concluding remarks are presented.

palabras clave/keywords:

Psicología basada en la evidencia, tests psicológicos, calidad de la investigación; generalización de la fiabilidad; meta-análisis; coeficiente de fiabilidad

Evidence-based psychology, psychological tests, research quality, reliability generalization, meta-analysis, reliability coefficient.

correspondencia: Laura.badenes@uv.es

Introducción

Los instrumentos de medida tales como los tests, escalas e inventarios (en adelante tests psicológicos) son de uso rutinario en la investigación y en la práctica profesional en Psicología y en las Ciencias de la Salud en general. De hecho, la investigación y la práctica profesional no son posibles si no es con la ayuda de los tests psicológicos que permitan evaluar el estado mental, psicológico, psicosocial, educativo o psiquiátrico de las personas. Estos instrumentos de medida resultan imprescindibles tanto en la prevención de trastornos psicológicos, como en su diagnóstico y en la valoración de los programas de intervención y tratamiento. Es por ello que evaluar la calidad métrica de los instrumentos de medida de uso habitual en Psicología y en las Ciencias de la Salud en general constituye un objetivo básico dentro del ámbito de la investigación y la práctica profesional.

Dentro de la calidad métrica de los instrumentos de medida, la fiabilidad se erige como una de las más importantes propiedades psicométricas cuando se aplican los test psicológicos a una muestra de participantes. La fiabilidad proporciona información sobre el grado de precisión de la medición asociada con una prueba (Slaney, 2017). La cantidad de error de medición determina la validez de las puntuaciones y, por lo tanto, el grado en que un instrumento de medida representa correctamente un constructo psicológico (Flake, Pek, y Hehman, 2017).

Dentro del marco de la Teoría Clásica de los Tests se han propuesto diferentes métodos de estimación de la fiabilidad de las puntuaciones de los tests psicológicos (e.g., Raykov y Marcoulides, 1971): (a) estabilidad temporal (e.g., fiabilidad test-retest), (b) consistencia interna (e.g., coeficiente alfa de Cronbach), (c) formas paralelas y (d) acuerdo entre evaluadores o intra-evaluadores. Cada uno de estos métodos estima la fiabilidad desde un punto de vista diferente. Así pues, la fiabilidad puede hacer referencia al grado en que los ítems del test son consistentes entre sí, o al grado en que las puntuaciones del test permanecen estables cuando éste se aplica a las mismas personas en diferentes ocasiones, o al grado en que diferentes evaluadores alcanzan valoraciones similares cuando aplican el test a los mismos individuos (Abad, Olea, Ponsoda, y García, 2011; Crocker y Algina, 1986; Traub, 1994).

En la práctica, los métodos de estimación de la fiabilidad basados en la consistencia interna y test-retest son los más utilizados (Flake et al., 2017). La consistencia interna evalúa el grado en que las partes (es decir, los ítems) o componentes (es decir, subconjuntos, mitades) de un test están correlacionadas, esto es, miden el mismo constructo. La medida de consistencia interna más utilizada es el coeficiente de alfa de Cronbach (Flake et al., 2017; Sánchez-Meca, López-Pina, y López-López, 2008), que se puede definir como la media de todas las posibles fiabilidades estimadas por el método de dos mitades (Cronbach, 1951).

El método test-retest captura la estabilidad temporal de las puntuaciones, es decir, hasta qué punto el rendimiento de una persona es repetible (Raykov y Marcoulides, 1971). La fiabilidad test-retest se define como la correlación entre las puntuaciones obtenidas en el mismo test, con el mismo grupo de sujetos, en dos ocasiones distintas, separadas por un intervalo de tiempo específico. Si las puntuaciones de los tests son fiables, las puntuaciones en la primera administración deben ser similares a las puntuaciones en la segunda administración

del test. Por lo tanto, se esperaría una alta correlación positiva entre las dos administraciones del test.

Sin embargo, la fiabilidad no es una propiedad inherente del test psicológico, sino más bien de las puntuaciones en una aplicación concreta del test, por lo que varía de una aplicación a otra de un test (Abad et al., 2011; Crocker y Algina, 1986; Raykov y Marcoulides, 1971). Algunos de los factores que afectan al coeficiente de fiabilidad son la longitud del test y la composición y variabilidad del grupo donde el test se aplica (Abad et al., 2011; Crocker y Algina, 1986; Raykov y Marcoulides, 1971). Así, cuanto mayor es el número de ítems que componen un test, mayor tenderá a ser el coeficiente de fiabilidad obtenido en una administración del mismo a una determinada muestra. Del mismo modo, cuanto más heterogéneo sea el grupo de participantes del estudio (mayor varianza de las puntuaciones observadas), mayor tenderá a ser el coeficiente de fiabilidad obtenido en una administración del mismo. También se puede dar diferente variabilidad en las puntuaciones de un test cuando éste se aplica a una muestra procedente de población general (no clínica), cuando se aplica a una muestra subclínica o cuando se administra a una muestra clínica. Por lo tanto, dado que la fiabilidad varía en función de la composición y la variabilidad de la muestra a la que se administra el test, los investigadores deberían examinar y reportar rutinariamente esta propiedad psicométrica con los datos de la muestra de su estudio (Slaney, 2017).

No obstante, bajo la creencia errónea de que la fiabilidad es una propiedad inherente al test, es común que los investigadores no informen de las estimaciones de fiabilidad obtenidas con los datos de la muestra, sino que las induzcan de aplicaciones previas del test (e.g., de los estudios de validación del test) o simplemente no informen de ninguna estimación de fiabilidad de las puntuaciones de los tests utilizados. Así, una reciente revisión sistemática realizada sobre 123 test psicológicos y más de 41,000 estudios empíricos mostró que el 78.6% de los estudios no informaron sobre las estimaciones de fiabilidad con los datos disponibles, sino que la indujeron de estudios previos (Sánchez-Meca, Rubio-Aparicio, López-Pina, Núñez-Núñez, y Marín-Martínez, 2015).

Esta práctica errónea de no informar de las estimaciones de fiabilidad obtenidas con los propios datos ha sido denominada ‹inducción de fiabilidad› (*reliability induction*) por Vacha-Haase, Kogan, y Thompson (2000). Además, la práctica de extrapolar la fiabilidad desde estudios previos está actualmente desaconsejada (Appelbaum et al., 2018; Thompson, 1994; Vacha-Haase et al., 2000; Wilkinson y the APA Task Force on Statistical Inference, 1999).

Afortunadamente, muchos investigadores no inducen la fiabilidad de las puntuaciones obtenidas con la administración de un test. Como consecuencia, es posible investigar cómo la composición y la variabilidad de las muestras utilizadas en diferentes aplicaciones de un determinado test psicológico pueden afectar la fiabilidad de las puntuaciones del test mediante la técnica metodológica de meta-análisis de generalización de la fiabilidad.

Meta-análisis de generalización de la fiabilidad

El meta-análisis (MA) es un tipo de revisión sistemática en la que se formula una pregunta con total claridad, se fijan unos criterios de selección de los estudios empíricos que abordan

esa pregunta, se lleva a cabo una búsqueda exhaustiva de esos estudios, se extraen datos relevantes de los estudios, se analizan estadísticamente y se alcanzan unas conclusiones sobre cuál es 'el estado del arte' sobre la pregunta en cuestión (Badenes-Ribera, 2016; Borenstein, Hedges, Higgins, y Rothstein, 2009; Botella y Gambara, 2002; Botella y Sánchez-Meca, 2015; Higgins y Green, 2008). En consecuencia, el MA es una herramienta metodológica que permite integrar cuantitativamente los resultados obtenidos a partir de un conjunto de investigaciones realizadas sobre una temática concreta. De este modo, los resultados de cada estudio individual, expresados en términos de tamaño del efecto, son combinados para obtener conclusiones más generales. Además, los estudios de meta-análisis permiten considerar aquellas variables que difieren entre los estudios primarios incluidos y que pueden contribuir a explicar la diferencia entre los resultados obtenidos en cada investigación particular (Botella y Sánchez-Meca, 2015).

El meta-análisis de generalización de la fiabilidad (MA GF) es un tipo de MA aplicado al ámbito psicométrico para integrar cuantitativamente alguna estimación de la fiabilidad obtenida al aplicar un determinado test psicológico (Vacha-Haase, 1998). Por lo tanto, el MA GF, dado que es un tipo específico de MA, también implica la recopilación de los estudios empíricos, en este caso, de los estudios que han aplicado un determinado test psicológico desde su creación, pero, a diferencia del MA, el resultado de cada estudio que se integra meta-analíticamente, no es un tamaño del efecto, sino alguna estimación de la fiabilidad obtenida al aplicar el test que se está analizando.

Utilidad de los meta-análisis de generalización de la fiabilidad

¿Qué puede ofrecernos un MA GF? Al aplicar técnicas estadísticas para integrar los resultados de un conjunto de estudios empíricos acerca de la fiabilidad de las puntuaciones de un test psicológico, un MA GF permite responder a preguntas tales como: (a) ¿Cuánto vale la fiabilidad media obtenida en las diversas aplicaciones de un mismo test? ¿cuál es la estimación por intervalo del coeficiente de fiabilidad promedio en la población?; (b) ¿Son homogéneos entre si los coeficientes de fiabilidad obtenidos en las diversas aplicaciones del test? O lo que es lo mismo, ¿se puede generalizar la fiabilidad de un determinado test a diferentes muestras y en diferentes contextos?; (c) ¿Son comparables los diferentes métodos de estimación de la fiabilidad de un test (consistencia interna, formas paralelas, test-retest, grado de acuerdo intercodificadores)?; y (d) Si los coeficientes de fiabilidad obtenidos en distintas aplicaciones de un test son muy heterogéneos, ¿cuáles son las características de las muestras, del contexto de aplicación y del propio test que pueden explicar tal heterogeneidad?.

En consecuencia, un MA GF permite a los investigadores caracterizar la fiabilidad promedio ponderada de las puntuaciones (con su significación estadística y su estimación por intervalo) obtenidas por un test psicológico en múltiples estudios y situaciones, y estimar el grado de variabilidad en los coeficientes de fiabilidad en diferentes tipos de medidas, muestras y contextos. Además, cuando los coeficientes de fiabilidad son heterogéneos, el MA GF permite explorar qué características de los estudios pueden estar estadísticamente relacionadas con las estimaciones de fiabilidad, esto es, explorar posibles variables moderadoras de los resultados (Botella, Suero, y Gambara, 2010; Rodriguez y Maeda, 2006; Sánchez-Meca,

López-López, y López-Pina, 2013; Sánchez-Meca y López-Pina, 2008; Sánchez-Meca, López-Pina, y López-López, 2009; Vacha-Haase, Henson, y Caruso, 2002; Vacha-Haase y Thompson, 2011). De esta manera, se puede determinar qué test psicológicos tienden a producir puntuaciones más fiables, para qué tipo de personas y para qué contextos.

Pero los MA GF, como cualquier otro MA o investigación primaria, no están exentos de sesgos, por lo que es fundamental saber hacer una lectura crítica de un trabajo de MA, siendo capaz de depurar y valorar su calidad metodológica. Una correcta lectura crítica de los resultados aportados por un MA GF sobre la fiabilidad de las puntuaciones de un test psicológico permite al lector valorar la fiabilidad de diferentes tests psicológicos que miden el mismo constructo psicológico y, en consecuencia, ayudarle en la toma de decisión sobre qué test psicológico aplicar en un caso particular. Así pues, en el ámbito profesional, los resultados ofrecerán datos objetivos y fiables sobre la precisión de los tests psicológicos habitualmente aplicados en la práctica profesional, lo que permitirá a los profesionales elegir con fundamento el mejor test para cada persona. Y en el contexto científico, los investigadores se beneficiarán de los resultados al disponer de datos objetivos y sistemáticos de cuál es la fiabilidad de los instrumentos de evaluación habitualmente utilizados en las investigaciones.

Sin embargo, la lectura crítica de los MAs GFs pasa necesariamente por que el investigador y el profesional de la Salud tenga unos conocimientos apropiados de qué es un MA GF, cómo se hace y a qué sesgos pueden estar expuestos sus resultados. Es por ello que este artículo se centra específicamente en cómo se hace y cómo se interpreta un MA GF.

Lectura crítica de un meta-análisis de generalización de la fiabilidad

En este apartado se explica las fases de un MA GF y se utiliza como ejemplo un reciente MA GF titulado "*A reliability generalization meta-analysis of self-report measures of muscle dysmorphia*" (Un meta-análisis de generalización de fiabilidad de las medidas de auto-informe de Dismorfia Muscular) elaborado por Rubio-Aparicio, Badenes-Ribera, Sánchez-Meca, Fabris, y Longobardi (2020) para ilustrar las claves principales en que los investigadores y profesionales de la Salud deben de fijarse cuando están leyendo un MA GF, con objeto de poder valorar críticamente la calidad de los resultados que aporta y su relevancia para la práctica clínica y profesional.

La realización de un MA GF implica seguir las mismas etapas que en cualquier MA (Botella y Sánchez-Meca, 2015; Sánchez-Meca et al., 2009). A saber (Badenes-Ribera, 2016, Botella y Sánchez-Meca, 2015; Rubio-Aparicio, Sánchez-Meca, Marín-Martínez, y López-López, 2018; Sánchez-Meca y Botella, 2010):

(1) Formulación del problema

(2) Selección de los estudios

(3) Codificación de los estudios

(4) Análisis estadístico e interpretación

(5) Publicación

(1) Formulación del problema. Se debe formular de forma clara y precisa la pregunta de investigación, lo que implica definir de forma teórica y operativa el constructo objeto de estudio y los instrumentos de medida que se han desarrollado para evaluarlos. Es importante tener en cuenta que un MA GF puede centrarse en valorar las estimaciones de fiabilidad de un test concreto, o también en analizar de forma comparativa las estimaciones de fiabilidad de diferentes tests elaborados para medir un mismo constructo. En el MA GF de Rubio-Aparicio et al. (2020) antes citado, se describen los siguientes objetivos, entre otros: estimar la fiabilidad promedio de las puntuaciones obtenidas por los diferentes tests desarrollados para evaluar la Dismorfia Muscular (en adelante, DM), examinar la variabilidad entre las estimaciones de fiabilidad; y buscar características sustantivas y metodológicas de los estudios que puedan asociarse estadísticamente con los coeficientes de fiabilidad. Además, en la introducción se define el trastorno de DM y se describen los tests psicológicos desarrollados para evaluar DM.

(2) Búsqueda de los estudios. Una vez determinada la pregunta de investigación, se establecen y definen los criterios de inclusión o selección de los estudios empíricos. Estos criterios deben incluir especificaciones relativas a las características de los participantes, al test psicológico utilizado en el estudio, y a las estimaciones de fiabilidad que se quieren meta-analizar. En el MA GF de Rubio-Aparicio et al. (2020), los estudios empíricos tenían que haber evaluado la DM utilizando alguno de los tests psicológicos desarrollados para medir sintomatología de DM y haber reportado alguna estimación de la fiabilidad calculada con sus propios datos. Respecto de las características de los participantes, se aceptaron los participantes provenientes de cualquier población (general, clínica, subclínica, etc.) y no se pusieron restricciones por edad.

Una vez establecidos los criterios de inclusión, se inicia propiamente la búsqueda de los estudios primarios, generalmente en bases electrónicas (e.g., WoS, PsycInfo, Medline, Scopus, ERIC, etc.) utilizando una combinación apropiada de palabras clave. También se suele consultar revistas especializadas en la temática en cuestión, y contactar con los autores relevantes en el tema de investigación para conseguir estudios de difícil localización, o estudios no publicados. El objetivo de utilizar diversas estrategias es que la búsqueda sea lo más comprehensiva posible para localizar el mayor número posible de estudios. Además, se recomienda presentar un diagrama de flujo que resuma el proceso de cribado (*screening*) y selección de los estudios.

En el MA GF realizado por Rubio-Aparicio et al. (2020), se consultaron las bases electrónicas WoS, Science Direct, PsycInfo, Medline (via Pubmed) usando las siguientes palabras clave en todos los campos del artículo: *muscle dysmorph**, muscle dysmorphia, reverse anorexia, bigorexia, vigorexia, y Adonis complex. Además, se consultaron las listas de referencias de los estudios seleccionados y de revisiones sistemáticas y MAs previos. Finalmente, para localizar estudios no publicados se contactó con autores reconocidos en el ámbito de la DM. La selección de los estudios se llevó a cabo de forma independiente por dos meta-analistas Doctores en Psicología, y las discrepancias entre ellos fueron resueltas por consenso.

En dicho MA GF se lograron seleccionar 61 estudios (N = 15,156 sujetos) que proporcionaron 73 estimaciones independientes de fiabilidad (coeficiente alfa de Cronbach y/o fiabilidad test-retest). La Figura 1 muestra el diagrama de flujo del proceso de selección de los estudios.

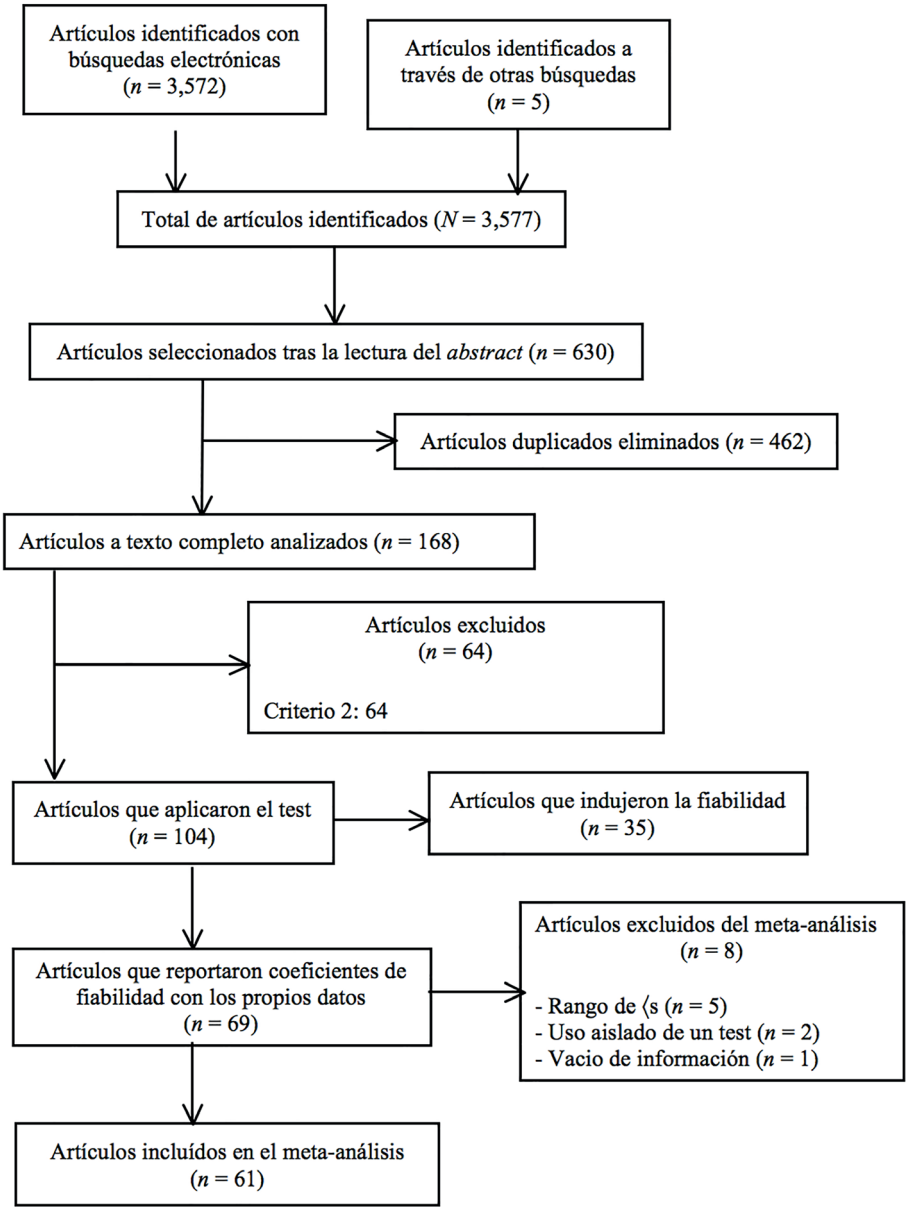

Figura 1.- Diagrama de flujo del proceso de selección adaptado de Rubio-et al. (2020).

(3) Codificación de los estudios. Supone registrar de manera estandarizada y sistemática las características de los estudios seleccionados. Se recomienda que la codificación de las variables se realice al menos por dos revisores de forma independiente y se evalúe e informe sobre la fiabilidad del proceso de codificación (al menos, los valores mínimos, máximos y media de los coeficientes de fiabilidad inter-jueces). Para ello, se elabora un Manual y un Protocolo de Codificación de las características relevantes de los estudios y que podrían actuar como moderadores de los resultados de fiabilidad analizados. En los MA GF se deben registrar características de los estudios tales como: media de las puntuaciones totales en el test y en las subescalas, la desviación típica de las puntuaciones totales en el test y en las subescalas, el número de ítems del test y de las subescalas, el intervalo de tiempo para la fiabilidad test-retest, el tamaño de la muestra y el tamaño de la muestra para la fiabilidad test-retest. Otras variables moderadoras que también se pueden registrar son la versión del test (original, adaptación, etc.), la edad de los sujetos, la distribución por sexo, etnia, nivel de estudios de la muestra, tipo de población (general, clínica, subclínica), etc.

En el MAGF de Rubio-Aparicio et al. (2020) se codificaron las siguientes variables: media de las puntuaciones totales en el test y en las subescalas, la desviación típica de las puntuaciones totales en el test y en las subescalas, el número de ítems del test y de las subescalas, intervalo de tiempo para la fiabilidad test-retest, el tamaño de la muestra, tamaño de la muestra para la fiabilidad test-retest, la versión del test (original, adaptación, etc.), edad media (y desviación típica) de los sujetos, la distribución por sexo (porcentaje de hombres), etnia (porcentaje de caucásicos), población diana I (general, gimnasios, estudiantes universitarios, clínica, subclínica, usuarios de esteroides anabólicos), población diana II (adultos, adolescentes o mixta), porcentaje de participantes clínicos en la muestra, tipo de trastorno clínico (DM vs otro trastorno, solo en muestras clínicas), localización geográfica del estudio, objetivo del estudio (estudio psicométrico vs estudio aplicado), diseño de la investigación, año del estudio, idioma del estudio, formación del investigador principal, y fuente de financiación.

Además, el proceso de codificación también se llevó a cabo por dos meta-analistas Doctores en Psicología de forma independiente y la fiabilidad Interjueces fue satisfactoria, con una correlación intraclase media de .96 (DT = 0.046, min. = .89, max. = 1) para variables continuas, y un coeficiente de Kappa medio de .90 (DT = .0087, min. = .80, max = 1) para las variables categóricas.

(4) Análisis estadístico e interpretación. El análisis estadístico involucra las estimaciones de los coeficientes de fiabilidad extraídas de los estudios empíricos y requiere determinar el modelo estadístico aplicado para integrar cuantitativamente las estimaciones de fiabilidad y para examinar el efecto de posibles variables moderadoras.

(4.1) Estimación de la fiabilidad. Una cuestión esencial en el MA GF es estimar la fiabilidad de las puntuaciones del test o de las subescalas. Los métodos de estimación de fiabilidad más utilizados en la práctica son el coeficiente de alfa de Cronbach, como medida de consistencia interna, y el coeficiente de correlación de Pearson, como medida de fiabilidad test-retest (Flake et al., 2017; Sánchez-Meca et al., 2008). Importante recordar que no se

deben mezclar los distintos coeficientes de fiabilidad obtenidos por distintos métodos de estimación, dado que los distintos coeficientes tienen distintas métricas y se refieren a fuentes de variación de diferente naturaleza (Botella y Sánchez-Meca, 215).

En esta fase, se extraen los coeficientes de fiabilidad reportados en los estudios seleccionados y calculados con sus propios datos. Se debe hacer una descripción clara de cómo se extrajeron y calcularon los coeficientes de fiabilidad de los estudios primarios. En este punto se debe decidir si los coeficientes de fiabilidad extraídos de los estudios deben ser transformados para normalizar su distribución y estabilizar sus varianzas (Botella y Suero, 2012). En el caso del coeficiente de alfa, una transformación adecuada es la propuesta por Hakstian y Whalen (1976) y recomendada por Rodríguez y Maeda (2006) y Sánchez-Meca et al. (2013). La transformación Hakstian-Whalen permite normalizar la distribución de los coeficientes de fiabilidad. Pero una mejor transformación de los coeficientes alfa es la propuesta por Bonett (2002) dado que permite normalizar la distribución de los coeficientes alfa y estabilizar sus varianzas (Sánchez-Meca et al., 2013). Y, para el caso de los coeficientes de fiabilidad test-retest, se recomienda utilizar la transformación de Z de Fisher (Sánchez-Meca et al. 2013).

Finalmente, como en el proceso de codificación de las características de los estudios, se recomienda que la extracción/cómputo de los coeficientes de fiabilidad se realice al menos por dos revisores de forma independiente y se evalúe e informe sobre la fiabilidad entre los evaluadores en el proceso de extracción/ estimación.

En el MA GF de Rubio-Aparicio et al. (2020) se extrajeron de los estudios primarios los coeficientes de alfa de Cronbach (consistencia interna) y los coeficientes de correlación de Pearson relativos a dos administraciones del mismo test en la misma muestra en dos momentos temporales distintos (fiabilidad test-retest). A estos coeficientes de fiabilidad se les aplicaron transformaciones para normalizar sus distribuciones y estabilizar sus varianzas. En concreto, los coeficientes alfa de Cronbach se transformaron por el método de Bonett (2002) y los coeficientes test-retest por el método de Z de Fisher. Después para facilitar la interpretación de los resultados del MA GF, los coeficientes de fiabilidad promedio y sus intervalos de confianza obtenidos con las transformaciones se devolvieron a las métricas del coeficiente alfa y de correlación de Pearson, respectivamente.

(4.2) Análisis estadístico e interpretación de los resultados. Una vez que se tienen registrados para cada estudio sus características (variables moderadoras) y sus coeficientes de fiabilidad, la base de datos resultante puede someterse a análisis estadísticos que permitan responder a las preguntas clave a que se enfrenta un MA GF. La primera pregunta clave es: *¿Cuál es el coeficiente de fiabilidad medio de los estudios?*. En esta fase, se debe especificar el marco estadístico (frecuentista o bayesiano) y el modelo estadístico meta-analítico asumido en los análisis (efectos fijos versus efectos aleatorios). El meta-analista debe explicar las razones para suponer un modelo de efectos fijos o aleatorios para analizar sus datos, dado que el modelo meta-analítico utilizado influye en los procedimientos estadísticos implementados para integrar cuantitativamente la información y en la generalización de los resultados (Hedges y Vevea, 1998). Estrechamente relacionado con la elección del modelo

estadístico está la decisión sobre si los coeficientes de fiabilidad deben ser ponderados en función de su precisión (o de algún otro factor) o no. Si se decide ponderar los coeficientes de fiabilidad según su precisión, caben al menos tres métodos de ponderación: ponderar en función del tamaño muestral y ponderar en función de la inversa de la varianza del coeficiente de fiabilidad que, a su vez, implica una fórmula diferente según que se asuma un modelo de efectos fijos o de efectos aleatorios (Bonett, 2002, 2010). En el MA GF de Rubio-Aparicio et al. (2020) desde un marco frecuentista, asumieron modelos de efectos aleatorios para acomodar la variabilidad exhibida por los diferentes coeficientes de fiabilidad, de forma que los coeficientes alfa de Cronbach se transformaron por el método de Bonett (2002) y los coeficientes test-retest por Z de Fisher. Después, para facilitar la interpretación de los resultados, los coeficientes de fiabilidad promedio y sus intervalos de confianza obtenidos con las transformaciones de Bonett o de Z de Fisher, se transformaron nuevamente a las métricas del coeficiente alfa y de la correlación de Pearson, respectivamente.

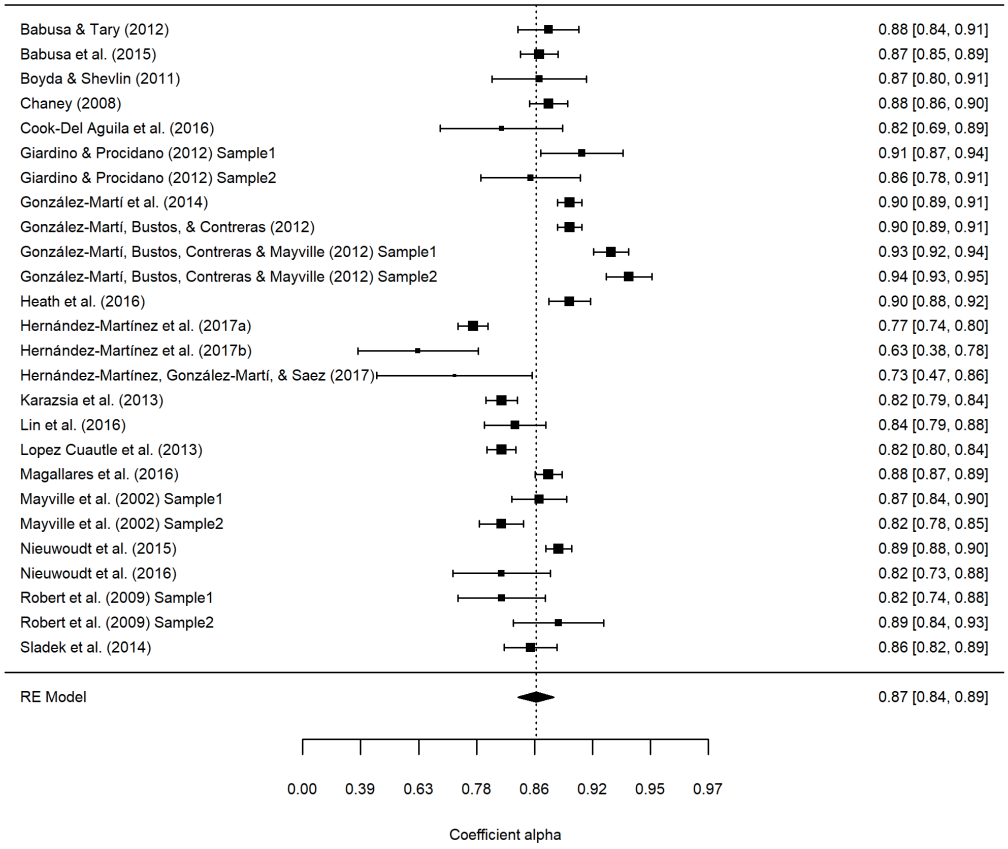

Figura 2.- Forest plot de los coeficientes alfa de Cronbach (Intervalo de confianza al 95%) para las puntuaciones totales del *Muscle Appearance Satisfaction Scale* (Tomado y adaptado de Rubio-Aparicio et al. 2020).

Los resultados del MA GF se suelen representar mediante la construcción de un gráfico denominado *Forest Plot* (Botella y Sánchez-Meca, 2015). En este gráfico se puede observar los coeficientes de fiabilidad obtenidos en cada estudio y sus intervalos de confianza, así como el coeficiente de fiabilidad medio y su intervalo de confianza. La Figura 2 muestra el *Forest Plot* del MA GF de Rubio-Aparicio et al. (2020) sobre los coeficientes de alfa de Cronbach para las puntuaciones totales del test "*Muscle Appearance Satisfaction Scale (MASS)*". En este estudio el coeficiente de fiabilidad medio fue de .87 (IC al 95%: .84, .89). La teoría psicométrica señala que la fiabilidad medida como consistencia interna debe ser superior a .80 para fines de investigación y superior a .90 para la práctica clínica (Nunnally y Bernstein, 1994). Sin embargo, los coeficientes alfa superiores a .70 pueden considerarse aceptables para la investigación exploratoria (Charter, 2003). Además, para evaluar la relevancia clínica de los coeficientes alfa, Cicchetti (1994) sugirió las siguientes pautas: inaceptable para coeficientes inferiores a .70, aceptable para el rango de .70 a .80, bueno para .80 a .90 y excelente para valores superiores a .90. Por lo tanto, los resultados de este MA GF para las puntuaciones totales en la escala MASS indican que esta escala muestra una fiabilidad media buena tanto para la investigación como para la práctica clínica.

Respecto de la segunda pregunta clave, *¿son homogéneos los coeficientes de fiabilidad de los estudios?*, la inspección visual del *Forest Plot,* permite examinar de forma exploratoria la similitud entre los coeficientes de fiabilidad de los estudios empíricos. Además, pruebas estadísticas específicas, como el estadístico Q de Cochran y el índice I^2, permiten tomar una decisión sobre si los coeficientes de fiabilidad son heterogéneos entre sí. En concreto, cuando el estadístico Q de Cochran alcanza la significación estadística ($p < .05$), ello es indicativo de heterogeneidad entre los coeficientes de fiabilidad (Higgins, Thompson, Deeks, y Altman, 2003). La información aportada por el estadístico Q se complementa con la del índice I^2, que cuantifica el grado de heterogeneidad, de forma que valores alrededor de 25%, 50%, y 75% denotan baja, moderada y alta heterogeneidad, respectivamente (Huedo-Medina, Sánchez-Meca, Marín-Martínez, y Botella, 2006).

En el caso del MA RG de Rubio-Aparicio et al. (2020), el *Forest Plot* muestra la heterogeneidad entre los coeficientes de fiabilidad de los estudios empíricos que oscilaron entre .63 y .94. Además, las pruebas estadísticas ejecutadas indicaron que los coeficientes de fiabilidad analizados eran muy heterogéneos entre sí ($Q = 35.63$, $p < .001$; $I^2 = 93.45$).

En el caso de no ser homogéneos los coeficientes de fiabilidad, se debe dar respuesta a la tercera cuestión clave: *¿Qué características de los estudios pueden dar cuenta de esa heterogeneidad observada?* En esta fase se analiza la influencia de variables moderadoras (categóricas y continuas) sobre los coeficientes de fiabilidad de los estudios, normalmente aplicando procedimientos de ponderación basados en los análisis de la varianza (ANOVA) para variables categóricas y en los modelos de regresión (meta-regresión) para variables continuas. La variable dependiente en esos análisis son los coeficientes de fiabilidad obtenidos en los estudios empíricos y las variables independientes o predictoras son las características de los estudios anteriormente codificadas. Se debe especificar el modelo estadístico meta-analítico asumido en los análisis de moderadores (efectos fijos versus efectos mixtos).

Se recomienda utilizar un modelo de efectos mixtos, ya que suele ser el más realista (Rubio-Aparicio, Sánchez-Meca, López-López, Marín-Martínez, y Botella, 2017).

En el análisis de variables moderadoras del MA GF de Rubio-Aparicio et al. (2020) se asumió el modelo de efectos mixtos para implementar tanto los ANOVAs ponderados para las variables moderadoras categóricas como los análisis de meta-regresión para las variables moderadoras continuas. En los ANOVAs, ninguna de las variables moderadoras alcanzó la significación estadística: versión del test aplicada (p = .858), foco del estudio (p = .347); población diana I (p = .842), población diana II (p = .491), lugar de realización del estudio (p = 691), idioma del estudio (p = .098), formación del investigador principal (p =.701), y fuente de financiación (p = .682). En los análisis de meta-regresión, nótese que el signo de la pendiente de regresión, bj, se obtuvo tomando los coeficientes alfa transformados mediante la fórmula de Bonett (2002) como variable dependiente. Dado que la transformación de Bonett invierte el sentido de la variable original (los coeficientes alfa), la dirección de la asociación verdadera entre los coeficientes alfa y la variable moderadora es la inversa de lo que está representado por el signo de la pendiente. En estos análisis sólo la edad media de los participantes (b = .04) estuvo relacionada con los coeficientes de fiabilidad. En concreto, mostró una relación negativa y estadísticamente significativa con las estimaciones de fiabilidad de las puntuaciones totales del MASS, lo que significa que las muestras con participantes más jóvenes exhibieron una mejor fiabilidad promedio que las muestras con participantes mayores. Este resultado sugiere que el MASS puede ser particularmente adecuado, no sólo como una medida general de DM, sino también en la realización de comparaciones entre grupos de participantes de diversos orígenes, excepto cuando los participantes tienen diferentes edades. El resto de variables moderadoras no presentaron relación con los coeficientes de fiabilidad: media de las puntuaciones totales (b = -.05, p = .611), desviación típica de las puntuaciones (b = -.05, p = .915), desviación típica de la edad (b = .03, p = 271), porcentaje de hombres (b = -.0001, p = .964), y porcentaje de caucásicos (b = -.0009, p =.889).

(4.3) Evaluación del sesgo de reporte. Los MA GF, como cualquier otro MA o investigación primaria, no están exentos de sesgos que pueden afectar a sus resultados. Uno de los principales sesgos al que se enfrenta el MA GF es el denominado 'sesgo de reporte' el cual puede afectar a la generalización de sus resultados. El sesgo de reporte se debe a la práctica de no reportar las estimaciones de fiabilidad obtenidas con los propios datos cuando la fiabilidad obtenida es baja (Appelbaum et al., 2018; Sánchez-Meca et al., 2015; Vacha-Haase et al., 2000). Es posible que en algunos estudios no se reporte el coeficiente de fiabilidad cuando éste ha alcanzado un valor bajo (e.g., por debajo de 0.7). No obstante, el hecho de que un estudio no reporte el coeficiente de fiabilidad no implica necesariamente que ésa haya sido la razón de su no reporte. Pueden darse otras razones del no reporte que no tienen por qué dar lugar a sesgos en las estimaciones (e.g., por olvido, por limitaciones de espacio en las revistas, por desconocimiento de la conveniencia de reportarlo, etc.). Si la razón de que los estudios no reporten el coeficiente de fiabilidad es por haber sido demasiado bajo, entonces se estaría incurriendo en un sesgo de reporte, de forma que los resultados del MA GF podrían ofrecer una sobreestimación de la fiabilidad exhibida por las puntuaciones del test. Dado que no podemos disponer de los coeficientes de fiabilidad de los estudios que no

lo reportan, un método indirecto, o aproximado, de evaluar si el no reporte del coeficiente de fiabilidad puede deberse a esta causa y, en consecuencia, que los resultados del MA GF puedan sufrir un sesgo al alza en la estimación de la fiabilidad, es comparar las características sociodemográficas de las muestras de los estudios que reportan las estimaciones de fiabilidad con las de los estudios que no las reportan (Sterne et al., 2011). Muy especialmente, la desviación típica de las puntuaciones del test es la característica fundamental que puede ayudar a comprobar si un MA GF puede estar sufriendo sesgo de reporte. Ello se debe a que, según la teoría psicométrica, la fiabilidad está positivamente relacionada con la variabilidad de las puntuaciones del test, de forma que, a mayor variabilidad, se espera una mayor fiabilidad (Crocker y Algina, 1986). En consecuencia, se puede comparar el promedio de las desviaciones típicas de los estudios que reportan la fiabilidad con el de los estudios que no la reportan. Si los estudios que no reportan la fiabilidad presentan un promedio de las desviaciones típicas significativamente menor que el de las que lo reportan, entonces sería plausible asumir que los estudios que no reportaron la fiabilidad presentaron coeficientes de fiabilidad más bajos que los de los estudios incluidos en el MA GF, dando lugar a sospecha de sesgo de reporte. Del mismo modo, se puede proceder comparando otras características sociodemográficas de los estudios que reportan o no la fiabilidad, tales como el promedio de las medias del test, el porcentaje promedio de varones (o mujeres) de las muestras, la distribución étnica (e.g., porcentaje de caucásicos) de las muestras, etc. Estas comparaciones deben hacerse por separado para los estudios que utilizaron muestras procedentes de población comunitaria, subclínica o clínica. En el análisis del sesgo de reporte del MA GF de Rubio-Aparicio et al. (2020), no se observaron diferencias significativas en las características de la muestra entre los estudios que indujeron y los que reportaron las estimaciones de fiabilidad con sus propios datos; pudiéndose descartar el sesgo de reporte, en términos de fiabilidad, como una amenaza para la validez de los resultados. Por lo tanto, sus resultados se pudieron generalizar razonablemente a todos los estudios que aplicaron el test MASS.

(5) Publicación. Como en cualquier otro estudio, la última fase de un estudio de MA de GF es su publicación para diseminar sus resultados. La redacción de un MA GF es muy similar a la de un MA, dado que es un tipo específico de MA. Las secciones de un artículo de MA RG suelen ser: introducción, método (criterios de inclusión, estrategia de búsqueda, extracción de datos, estimación de la fiabilidad, análisis estadísticos), resultados, discusión y conclusiones (e.g., Rubio-Aparicio et al., 2020). Recientemente Sánchez-Meca et al. (2017, Julio) han propuesto una guía para el correcto reporte de MAs de GF: la guía REGEMA (*REliability GEneralization Meta-Analysis)*. Esta guía se puede utilizar para redactar o para valorar críticamente un MA de GF. La guía está orientada a verificar si los meta-analistas han hecho explícitas todas las decisiones que han tomado durante la realización del MA de GF, lo cual es fundamental para poder valorar su calidad de manera crítica y para garantizar que otros investigadores puedan replicar el estudio MA de GF.

Consideraciones finales

Los resultados de un MA GF permiten ofrecer pautas a los investigadores y profesionales de la salud sobre qué tests psicológicos son más fiables para evaluar un determinado constructo

y en qué circunstancias. Pero los MA GF, como cualquier otro MA o investigación primaria, no están exentos de sesgos, por lo que es fundamental saber hacer una lectura crítica de un trabajo de MA, siendo capaz de depurar y valorar su calidad metodológica.

La lectura crítica de los MA GF implica que el lector tenga unos conocimientos previos apropiados de qué es un MA GF, cómo se hace y a qué sesgos pueden estar expuestos sus resultados. Sólo la comprensión de todas las fases por la que discurre un MA GF y de las decisiones tomadas por los meta-analistas en la realización del mismo permitirá a los lectores hacer una evaluación crítica de los resultados. Conscientes de esta problemática, se ha elaborado este artículo centrado específicamente en cómo se hace y cómo se interpreta un MA GF para intentar ayudar a los psicólogos en la toma de sus decisiones, tanto en la investigación aplicada como en su ejercicio profesional.

Agradecimientos

Este trabajo fue apoyado con fondos del Ministerio de Economía y Competitividad del Gobierno de España y por el Fondo Europeo de Desarrollo Regional (FEDER) Proyecto No. PSI2016-77676-P.

Referencias

Abad, F. J., Olea, J., Ponsoda, V. y García, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Síntesis.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., y Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist, 73*, 3–25.

Badenes-Ribera, L. (2016). *Tamaño del efecto y su intervalo de confianza y meta-análisis en Psicología*. Tesis Doctoral. Universitat de València.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*(4), 335–340. doi: 10.3102/10769 98602 7004335

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

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

Botella, J., y Gambara, H. (2002). ¿Qué es el meta-análisis? Madrid: Biblioteca Nueva.

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

Botella, J., y Suero, M. (2012). Managing heterogeneity of variances in studies of internal consistency generalization. *Methodology, 8*, 71-80.

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

Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability methods, and the clinical implications of low reliability. *The Journal of General Psychology, 130,* 290–304. doi: 10.1080/00221300309601160

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessments instruments in psychology. *Psychological Assessment, 6,* 284–290. doi:10.1037/1040-3590.6.4.28

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334. doi: 10.1007/BF02310555

Flake, J. K., Pek, J., y Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*, 370–378. doi: 10.1177/1948550617693063

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

Hedges, L. V., y Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504. doi: 10.1037/1082-989X.3.4.486

Higgins, J. P. T. y Green, S. (Eds.) (2008). *Cochrane handbook for systematic reviews of interventions*. Chichester, UK: Wiley-Blackwell.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., y Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557–560. doi: 10.1136/ bmj.327.7414.557

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., y Botella, J. (2006). Assessing heterogeneity in meta-analysis: *Q* statistic or I^2 index? *Psychological Methods, 11,* 193-206. doi: 10.1037/1082-989X.11.2.193

Nunnally, J. C., y Bernstein, I. H. (1994). *Psychometric Theory*. New York, NY: McGraw Hill.

Raykov, T., y Marcoulides, G. A. (1971). *Introduction to psychometric theory*. New York, NY: Taylor and Francis Group.

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

Rubio-Aparicio, M., Badenes-Ribera, L., Sánchez-Meca, J., Fabris, M. A., y Longobardi, C. (2020). A reliability generalization meta-analysis of self-report measures of muscle dysmorphia. *Clinical Psychology: Science and Practice, 27*, e12303. doi: 10.1111/cpsp.12303

Rubio-Aparicio, M., Sánchez-Meca, J., López-López, J.A., Marín-Martínez, F., y Botella, J. (2017). Analysis of categorical moderators in mixed-effects meta-analysis: Consequences of using pooled versus separate estimates of the residual between-studies variances. *British Journal of Mathematical and Statistical Psychology, 70*, 439–456. doi: 10.1111/bmsp.12092

Rubio-Aparicio M., Sánchez-Meca, J., Marín-Martínez, F., y López-López, J. A. (2018). Guidelines for Reporting Systematic Reviews and Meta-analyses. *Anales de Psicologia, 34,* 412-420. doi: 10.6018/analesps.34.2.320131

Sánchez-Meca, J., y Botella, J. (2010). Revisiones sistemáticas y Meta-análisis: Herramientas para la práctica profesional. *Papeles del Psicólogo, 31,* 7-17.

Sánchez-Meca, J., y López-López, J. A. y López-Pina, J. A. (2013). Some recommended statistical analytic practices when reliability generalization studies are conducted. *British Journal of Mathematical and Statistical Psychology, 66*, 402-425.

Sánchez-Meca, J., y López-Pina, J. A. (2008). El enfoque meta-analítico de generalización de la fiabilidad. *Acción Psicológica, 5*, 37-64.

Sánchez-Meca, J., López-Pina, J. A., y López-López, J. A. (2009). Generalización de la fiabilidad: Un enfoque meta-analítico aplicado a la fiabilidad. *Fisioterapia, 31*, 262-270.

Sánchez-Meca, J., López-Pina, J. A., Rubio-Aparicio, M., Marín-Martínez, F., Núñez-Núñez, R. M., López-García, J. J., y López-López, J. A. (2017, Julio). *REGEMA: Propuesta de una guía para la realización y reporte de meta-análisis de generalización de la fiabilidad*. Comunicación presentada en el XV Congreso de Metodología de las Ciencias Sociales y de la Salud, Barcelona (España).

Sánchez-Meca, J., Rubio-Aparicio, M., López-Pina, J. A., Núñez-Núñez, R. M. y Marín-Martínez, F. (2015, Julio). *El fenómeno de la inducción de la fiabilidad en ciencias sociales y de la salud.* Comunicación presentada en el XIV Congreso de Metodología de las Ciencias Sociales y de la Salud (Palma de Mallorca).

Slaney, K. (2017). *Validating psychological constructs*. UK: Palgrave Macmillan. doi: 10.1057/978-1-137-38523-9

Sterne, J. A., Sutton, A. J., Ioannidis, J., Terrin, N., Jones, D. R., Lau, J., …, & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal, 343,* D4002. doi: 10.1136/bmj.d4002

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*, 837-847.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications* (Vol. 3). Thousand Oaks, CA: Sage.

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

Vacha-Haase, T., y Henson, R. K., y Caruso, J. C. (2002). Reliability Generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement, 62*, 562–569. doi: 10.1177/0013164402062004002

Vacha-Haase, T., Kogan, L. R., y Thompson, B. (2000). Sample compositions and variabilities in published studies versus those of test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60,* 509–522. doi: 10.1177/ 00131640021970682

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

Wilkinson, L., y Task Force on Statistical Inference, American Psychological Association, Science Directorate. (1999). Statistical methods in psychology journal: Guidelines and explanations. *American Psychologist, 54,* 594–604. doi: 10.1037/0003-066X.54.8.594