Fisioterapia 2009;**31(6)**:262–270

Fisioterapia

www.elsevier.es/ft

REVISIÓN

Generalización de la fiabilidad: un enfoque metaanalítico aplicado a la fiabilidad

J. Sánchez-Meca*, **J.A. López-Pina y J.A. López López**

Departamento de Psicología Básica y Metodología, Facultad de Psicología, Campus de Espinardo, Universidad de Murcia, Murcia, España

Recibido el 20 de octubre de 2008; aceptado el 14 de mayo de 2009
Disponible en Internet el 22 de octubre de 2009

PALABRAS CLAVE
Generalización de la fiabilidad;
Metaanálisis;
Coeficiente de fiabilidad

Resumen
La fiabilidad no es una propiedad inherente al test, por lo que frases del tipo ''la fiabilidad del test es de 0,80'' son incorrectas. Ello se debe a que la fiabilidad es una propiedad de las puntuaciones obtenidas por un test en una aplicación concreta de éste. La generalización de la fiabilidad (GF) es un nuevo tipo de metaanálisis que permite examinar empíricamente la variabilidad de las estimaciones de la fiabilidad en diferentes aplicaciones de un test. Los estudios de GF están poniendo en evidencia lo inadecuado que resulta esa práctica habitual de los investigadores de inducir la fiabilidad a partir de estimaciones previas de ésta. En este artículo se presenta una panorámica del enfoque de GF, describiendo cuáles son sus fases de realización. Además, se discuten algunos de los problemas estadísticos más importantes de los estudios GF, tales como: a) procedimientos de transformación de los coeficientes de fiabilidad; b) métodos de ponderación de los coeficientes, y c) modelos estadísticos asumibles.

KEYWORDS
Reliability generalization;
Meta-analysis;
Reliability coefficient

Reliability generalization: A meta-analytic approach to reliability coefficients

Abstract
Reliability is not a property inherent to the test, so that sentences such as ''the test reliability is 0.80'' are wrong. That is because reliability is a property of scores obtained in a given application o a test. Reliability generalization (RG) is a new kind of meta-analysis which enables to empirically examine the variability of the reliability estimates across different applications of a test. The RG studies are evidencing how unadvisable is the usual practice of researchers of inducing reliability from previous estimates. In this article an overview of the RG approach is presented, describing the required steps. Moreover, some

*Autor para correspondencia.
Correo electrónico: jsmeca@um.es (J. Sánchez-Meca).

0211-5638/$ - see front matter © 2008 Asociación Española de Fisioterapéutas. Publicado por Elsevier España, S.L. Todos los derechos reservados.
doi:10.1016/j.ft.2009.05.005

of the most important statistical issues concerning RG studies are discussed, such as: (a) transforming procedures of the reliability coefficients, (b) weighting methods of the coefficients, and (c) statistical models that can be assumed.

La medida es el componente más importante dentro del proceso de una investigación[1]. Tanto la investigación como la práctica profesional en ciencias de la salud, en general, y en fisioterapia, en particular, está plagada del uso y aplicación de test e instrumentos baremados psicométricamente, con el propósito de medir en las personas, y en especial en los pacientes, el nivel que éstos tienen de variables médicas, psicológicas y de salud en general, que tienen un gran impacto en la calidad de vida de los seres humanos. Conocer la calidad métrica de los instrumentos de medida constituye, pues, una tarea esencial para el profesional de las ciencias de la salud. Una de las labores más relevantes que debe realizar el profesional es calcular, o conocer, la fiabilidad de las puntuaciones de un test.

En la práctica, no es habitual que las investigaciones que han empleado un test aporten estimaciones para la fiabilidad de los datos de la muestra utilizada. Lo que sí aparece a menudo son alusiones a los valores de la fiabilidad obtenidos en una aplicación previa del instrumento. Esto tendría sentido si la fiabilidad fuese un valor inmutable del test a lo largo de diferentes aplicaciones. Sin embargo, se trata de una propiedad referida a los resultados obtenidos con un instrumento de medida, y no al instrumento en sí[2–4].

La concepción estática de la fiabilidad (es decir, considerarla como un valor fijo para un test) está muy asentada en parte de la comunidad científica, y se habla de ella como propiedad del instrumento con demasiada frecuencia en los artículos de investigación. Un test, sin embargo, no es en sí más o menos fiable. Por ejemplo, puede producir puntuaciones altamente consistentes en una aplicación y, en ocasiones posteriores, generar datos a partir de los cuales la estimación de la fiabilidad tenga un valor mucho más bajo[5]. Aunque las muestras de sujetos a los que se administró el test pertenezcan a una misma población, las sucesivas aplicaciones variarán debido al azar derivado del proceso de muestreo. Pero si, además, las muestras seleccionadas para cada aplicación proceden de poblaciones diferentes, entonces la variación de los datos será mayor a la que cabría esperar por mero error de muestreo aleatorio[6,7].

De acuerdo con lo anterior, aludir a la fiabilidad reportada en investigaciones previas sólo sería apropiado si la muestra actual fuera igual en composición y variabilidad a la anterior. Esto contrasta con la frecuencia con la que los investigadores hacen referencia a las estimaciones de la fiabilidad de aplicaciones pretéritas del test, especialmente el manual (donde se habrá llevado a cabo un primer estudio de sus propiedades psicométricas fundamentales). Esta práctica de asumir para una determinada muestra alguna estimación previa de la fiabilidad en otra muestra, ha sido denominada por Vacha-Haase[7] como inducción de la fiabilidad (*reliability induction*). Hablamos de inducción porque el investigador parte de un caso particular (la estimación obtenida en una administración anterior del instrumento), y lo extiende, como si fuera generalizable, a los datos de su propia muestra.

En la última década se han llevado a cabo numerosos estudios que dejan al descubierto este problema en la literatura científica. Por ejemplo, Vacha-Haase y otros[8] revisaron las prácticas en el reporte de la fiabilidad de todos los artículos de investigación publicados en tres revistas psicológicas (Journal of Counseling Psychology, Psychology & Aging y Professional Psychology) entre los años 1990 y 1997, con un volumen total de 839 artículos. Sus análisis mostraron que sólo un 35,6% proporcionaba coeficientes de fiabilidad para los datos del estudio, mientras que un 22,9% la inducía de estudios previos, un 3,8% hacía alusión a la fiabilidad del instrumento en estudios anteriores sin valores concretos y, por último, un 36,4% ni siquiera mencionaba el concepto de fiabilidad. En esta misma línea, Whittington[9] encontró en su revisión de estudios publicados en 22 revistas del ámbito de la educación que el 54% de éstos indujeron la fiabilidad desde otras aplicaciones de los test. Y Vacha-Haase et al[10], en su revisión de 25 estudios de generalización de la fiabilidad (GF) encontraron que, en promedio, el 75,6% de los estudios empíricos que utilizan instrumentos de medida indujeron la fiabilidad a partir de anteriores administraciones del instrumento, mientras que sólo el 25,2% de los estudios aportan estimaciones propias de la fiabilidad.

Para que este proceso de inducción de la fiabilidad tuviera cierta validez, los investigadores tendrían que comprobar que su grupo es similar en composición y variabilidad a las del grupo del estudio en el que se calculó el coeficiente de fiabilidad. Sin embargo, no siempre es posible efectuar esta comprobación y, en cualquier caso, se realiza en muy raras ocasiones.

La fiabilidad se refiere a la consistencia o replicabilidad de las puntuaciones, siendo con ello un reflejo de la calidad de la medida del instrumento en una aplicación concreta. De manera lógica, una baja fiabilidad atenúa la estimación del tamaño del efecto y disminuye la potencia estadística de las pruebas de significación[11]. Por ello, la validez de las conclusiones estará sujeta necesariamente a la fiabilidad del instrumento que se ha empleado en la fase de evaluación[12]. En cuanto a los artículos que reportan la fiabilidad de estudios previos, esto es mejor que nada, ya que al menos denota una conciencia del hecho de que un valor bajo de la fiabilidad atenúa el tamaño del efecto. Sin embargo, los investigadores deberían considerar los factores bajo los cuales resulta factible inducir la fiabilidad[7], a los que ya hemos hecho alusión.

El énfasis de la comunidad científica en el adecuado reporte de la fiabilidad por parte de los investigadores ha ido en aumento a lo largo de los últimos años. Quizá la

recomendación más reseñable al respecto fue la promulgada en 1999 por la APA Task Force on Statistical Inference[11], donde se afirmaba en la página 596 que "la fiabilidad es una propiedad de las puntuaciones de un test para una muestra particular de sujetos", concluyendo más adelante que "los autores deberían proporcionar los coeficientes de fiabilidad para los datos que se están analizando, incluso cuando el foco de su investigación no sea psicométrico". Recomendaciones similares se han propuesto desde otras importantes asociaciones científicas, tales como la American Educational Research Association y el National Research Council on Measurement in Education. Este cambio de mentalidad también se ha visto reflejado en las políticas editoriales de algunas revistas, tales como Educational and Psychological Measurement[13] o Journal of Experimental Education[14], que han incluido el reporte adecuado de la fiabilidad entre los requisitos para los autores que deseen publicar sus investigaciones.

El enfoque de generalización de la fiabilidad

Dado que la fiabilidad es una propiedad de las puntuaciones y no del test psicométrico, para cada aplicación del test se podrán determinar uno o más coeficientes de fiabilidad que podrán variar en función de diversos factores (errores de muestreo, modo y condición de aplicación del test, composición y variabilidad de la muestra). Por ello, estudiar cómo varían los coeficientes de fiabilidad en cada grupo, sea normativo o no, constituye una tarea científica que el investigador no puede eludir.

Para abordar esta tarea, la metodología idónea es el metaanálisis, ya que permite integrar cuantitativamente los resultados numéricos de un conjunto de estudios sobre un mismo tema, aplicando para ello las mismas normas de rigor científico que se exigen a los estudios empíricos[15–24]. Aplicado al estudio de la fiabilidad de las puntuaciones, el metaanálisis permite integrar mediante técnicas de análisis estadístico, los coeficientes de fiabilidad que se obtengan a partir de la aplicación de un test a grupos con distintas características. Esta integración permite obtener una estimación de la fiabilidad media de las puntuaciones, estudiar la variabilidad de los coeficientes de fiabilidad y si tal variabilidad es muy elevada (más de la esperable por puro error de muestreo aleatorio), tratar de identificar qué características de los estudios pueden estar provocándola.

Las utilidades del enfoque de GF son diversas. Su aparición constituye en parte una crítica a las prácticas erróneas de numerosos investigadores, a las que subyace una concepción equivocada del concepto de fiabilidad y sus implicaciones. Así, este enfoque nace como un instrumento para denunciar un error frecuente en la literatura científica, y para clarificar el concepto de fiabilidad y, con ello, alcanzar una mayor tasa de reporte de la fiabilidad de las puntuaciones en los trabajos en los que se haya utilizado un test. Por otra parte, los resultados de un estudio de GF interesan directamente a los expertos en medición, ya que ayudan a una mejor comprensión de los factores que influyen en el coeficiente de fiabilidad de las puntuaciones tras una aplicación del test. Por último, tampoco podemos olvidar el valor que las conclusiones de estos estudios suponen a los futuros usuarios del test en el ámbito aplicado.

El enfoque de la GF también tiene sus detractores[25–27], pero desde su inicio en 1998 hasta la fecha ya se han contabilizado más de 50 estudios publicados. Entre los más importantes, en cuanto a la escala objeto de análisis, encontramos los realizados sobre el Beck Depression Inventory[28], el Spielberger State-Trait Anxiety Inventory[29], la Psychopathy Checklist[30], el Balanced Inventory of Desirable Responding[31], o las escalas de locus de control de Rotter y de Nowicki-Strickland[32].

A continuación, presentamos una revisión de los procedimientos más usuales para estimar la fiabilidad. Seguidamente desarrollaremos las fases que comporta un estudio de GF, deteniéndonos en los aspectos analíticos y estadísticos de esta metodología.

Métodos para estimar la fiabilidad

Para estimar la fiabilidad de las puntuaciones de un grupo se pueden emplear uno o más de estos métodos: test-retest, formas paralelas y dos mitades[2]. El método test-retest requiere la aplicación del test en dos ocasiones diferentes. El método de las formas paralelas supone aplicar dos formas del mismo test estrictamente paralelas (iguales medias y varianzas). Por último, el método de las dos mitades consiste en dividir el test en dos partes iguales, de modo que sólo una aplicación del test permite obtener una estimación de la fiabilidad de las puntuaciones. Los tres métodos dependen del coeficiente de correlación de Pearson para obtener una evaluación empírica de la fiabilidad de las puntuaciones, pero ya que en el procedimiento de las dos mitades se divide el test en dos partes equivalentes, el coeficiente de fiabilidad obtenido es el coeficiente del test mitad, por lo que se requiere emplear la ecuación de Spearman-Brown para el caso de longitud doble para obtener el coeficiente de fiabilidad en el test completo[2].

El procedimiento de las dos mitades es el de uso más extendido en la práctica actual para evaluar la fiabilidad de las puntuaciones, en especial una de sus variantes: el coeficiente alfa[33]. Este estimador asume que si el test tiene j ítems, en realidad tiene j mitades, por lo que la obtención del coeficiente de fiabilidad se realiza mediante el promedio de las $j(j-1)$ covarianzas entre los ítems del test, suponiendo que los ítems tienen iguales varianzas de error. La expresión del coeficiente alfa es:

$$\alpha \leq \frac{J}{J-1}\left(1-\frac{\sum \sigma_j^2}{\sigma_X^2}\right),$$

donde J es el número de ítems, ${\sigma_j}^2$ es la varianza del ítem j, y ${\sigma_X}^2$ es la varianza total.

Fases de un estudio de generalización de la fiabilidad

Dado que un estudio de GF es un tipo de metaanálisis, sus etapas son básicamente las mismas que las que se suelen proponer para los metaanálisis: 1) formulación del problema; 2) búsqueda de los estudios; 3) codificación de los

estudios; 4) análisis estadístico e interpretación, y 5) publicación del estudio[15,34–37].

Formulación del problema

En un estudio de GF, el objetivo fundamental es el de examinar la variabilidad de las estimaciones de la fiabilidad obtenidas al aplicar un test en diferentes contextos y a diferentes grupos, que pueden proceder de diferentes poblaciones de referencia. Una tarea importante para alcanzar esta meta consiste en identificar las características de los estudios que afectan a los coeficientes de fiabilidad obtenidos en las aplicaciones del instrumento.

El test se puede haber aplicado en distintos contextos, con diferentes fines o propósitos (p. ej., diagnóstico de un trastorno, cribado de población general, etc.), pueden haber varias versiones del test (p. ej., una versión más corta respecto de la original), o puede haberse traducido y/o adaptado a diferentes idiomas y/o culturas, o también a diferentes edades. Todos estos factores pueden afectar a la fiabilidad de las puntuaciones del test y justificarían la conveniencia de llevar a cabo un estudio de GF de las puntuaciones obtenidas con el test.

La decisión sobre si un estudio de GF es apropiado para un test concreto puede depender de dos aspectos[6]. Por un lado, el test debe estar suficientemente extendido en la comunidad científica y no ser demasiado reciente para que tenga sentido integrar las estimaciones de la fiabilidad.

Por otra parte, debería existir un número de estudios empíricos suficiente que aporten estimaciones propias de la fiabilidad de las puntuaciones, así como otros datos estadísticos relevantes, en especial la variabilidad de las puntuaciones en la muestra. En cuanto a los coeficientes, no es posible indicar un número mínimo como criterio para decidir si es apropiado o no realizar un estudio de GF. Los estudios de GF realizados hasta la fecha son muy variables a este respecto. Así, el número de estimaciones de la fiabilidad metaanalizadas puede ser tan bajo como los 18 coeficientes alfa integrados en el estudio de Campbell[30] sobre el test Psychopathy Checklist, o tan elevado como los 813 coeficientes alfa del estudio de Leach[38] sobre el test Self-Description Questionnaire.

Búsqueda de los estudios

En esta etapa, el primer paso consiste en definir claramente los criterios de selección de los estudios, entre los que podemos destacar los siguientes: a) los estudios seleccionados tienen que ser empíricos y grupales, es decir, tienen que haber utilizado uno o varios grupos donde se ha aplicado el test objetivo; b) si existen diferentes versiones de diferente longitud, o bien existen diferentes adaptaciones idiomáticas, culturales o de edades, tenemos que especificar si nuestro estudio de GF se centrará en la escala original únicamente o si, por el contrario, interesa examinar todo el conjunto de diferentes versiones que a lo largo de la vida del test se pueden haber desarrollado; c) también es necesario referenciar la población sobre la que se realizará el estudio de GF, ya que las puntuaciones del test pueden no tener la misma fiabilidad en aplicaciones clínicas o en grupos procedentes de la población normal, o cuando se aplica el test a franjas de edad diferentes; d) es preciso especificar el idioma en el que tiene que estar escrito el trabajo, ya que las limitaciones propias del equipo de investigación impedirán la inclusión de estudios escritos en aquellos idiomas que dicho equipo no domine, y e) por último, es preciso determinar el período temporal de la búsqueda: año de inicio, que será generalmente la fecha de construcción del test, y año final de la búsqueda.

Al menos, todos estos aspectos deberán tenerse en cuenta en la definición de los criterios de selección de los estudios, pero dependiendo del instrumento de medida en cuestión es posible que sea necesario incorporar otros criterios de selección adicionales.

Una vez que se han fijado los criterios de selección de los estudios, se debe diseñar un plan de búsqueda combinando diferentes sistemas. Una primera aproximación al número de estudios se puede obtener de las bases de datos electrónicas (PsycInfo, Medline, ERIC, etc.) a través del *abstract* o resumen de los estudios, aunque también deben consultarse otras bases de datos, tales como las de las colaboraciones Cochrane (www.cochrane.org) y Campbell (www.campbellcollaboration.org), que son dos asociaciones internacionales dirigidas a promover la realización de estudios metaanalíticos de alta calidad en el ámbito de la salud, la educación, el trabajo social y la criminología[39–41]. Como complemento a estas estrategias de búsqueda se puede recurrir al buscador Google académico y utilizar el mismo criterio que en la búsqueda anterior: que figure el nombre del test en el *abstract* del documento. Con este procedimiento de búsqueda podremos identificar estudios que han utilizado el test y que no fueron detectados por la estrategia anterior.

No obstante, es probable que el test haya sido empleado en algunos estudios, se haya obtenido su fiabilidad pero no se informe de su aplicación en el *abstract* del trabajo, así que será necesario recurrir a otras estrategias de búsqueda complementarias, algo más informales, que consisten en: a) examinar estudios de revisión en los que sabemos que se ha incluido alguna referencia al test; b) revisar estudios metaanalíticos sobre temas que tienen que ver con el test referenciado, y c) consultar a investigadores expertos en el tema para que nos envíen trabajos en los que se ha aplicado el test. Estas estrategias informales pueden ayudar a localizar estudios no publicados y de difícil localización por no estar recogidos en los repertorios ni en las bases internacionales.

Una vez que se han localizado los estudios que han realizado alguna aplicación del test debemos comprobar si han obtenido alguna estimación de la fiabilidad de las puntuaciones con los datos del/de los grupo/s empleado/s. El conjunto final de trabajos incluidos en nuestro estudio de GF estará formado por aquellos estudios empíricos que hayan aplicado el test y aporten al menos una estimación de la fiabilidad con los datos de la propia muestra de sujetos.

Codificación de los estudios

Habitualmente, los valores de los coeficientes de fiabilidad procedentes de los distintos estudios difieren notablemente entre sí. Uno de los objetivos del investigador que acomete la elaboración de un estudio de GF, consiste en buscar

factores que puedan explicar parte de la variabilidad entre las estimaciones recuperadas. La fase de codificación es el proceso mediante el cual registramos las características de los estudios que podrían explicarnos parte de esta variabilidad en los coeficientes de fiabilidad.

Para dilucidar qué variables pueden resultarnos útiles, podemos guiarnos por las indicaciones de la teoría psicométrica[2], así como por estudios de GF previos. En cualquier caso, una idea prudente es la de codificar suficientes variables como para que se puedan llevar a cabo numerosos análisis estadísticos, incluso aunque los estudios apenas aporten información sobre alguna de las características que en principio parecían más decisivas[6].

Un primer grupo de variables que pueden estar afectando al valor de los coeficientes son los factores metodológicos, como las diferentes formas de aplicación del test (autoinforme vs. aplicación por un evaluador), diferentes formatos de recogida de las respuestas (respuestas en papel y lápiz vs. informatizadas), distintas versiones del test (versión larga vs. corta del test), diferentes adaptaciones del test a otros idiomas, culturas (versión original del test vs. versiones adaptadas) o edades (niños, adolescentes, adultos, tercera edad), el tamaño del grupo y la variabilidad de las puntuaciones del test en el grupo. Otro conjunto de factores que se deben considerar tiene que ver con la procedencia del grupo, su composición y la población de referencia. Dentro de esta categoría se pueden citar la naturaleza clínica vs. normal de la población de referencia, la edad de los sujetos de la muestra (y su variabilidad), así como la distribución por sexo, por etnia, por nivel educativo, por estatus socioeconómico, etc.

Un tercer conjunto de características que también pueden provocar variabilidad en los coeficientes de fiabilidad de un mismo test son de tipo contextual, como por ejemplo, el propósito del estudio, distinguiendo entre estudios psicométricos (p. ej., estudio de validación de un test, adaptación de un test, etc.) y estudios de naturaleza sustantiva (p. ej., estudio predictivo de factores de riesgo de un trastorno, sobre la eficacia de un tratamiento, estudios diagnósticos, etc.). También pueden tener un efecto contextual el país o el continente en el que se realizó el estudio, el año de realización o de publicación del estudio, el criterio diagnóstico utilizado cuando se trata de población clínica, etc.

Un aspecto muy importante en un estudio de GF es obtener estimaciones de la fiabilidad con los datos propios del grupo. Así, si el estudio incluye más de un coeficiente de fiabilidad para el grupo, se deben recoger todos, por lo que más que el estudio (o el artículo), la unidad de análisis en un estudio de GF es el grupo, de ahí que un mismo estudio pueda aportar al metaanálisis más de una unidad de análisis.

En segundo lugar, también es posible que un estudio presente más de una estimación de la fiabilidad sobre un mismo grupo. Por ejemplo, el estudio puede haber calculado el coeficiente alfa y el coeficiente de fiabilidad test-retest sobre las puntuaciones de una misma muestra. En este caso, también debemos recoger ambas estimaciones de la fiabilidad, si bien se metaanalizarán por separado para evitar problemas de dependencia estadística. El protocolo de registro de cada estudio debe, pues, contemplar la posibilidad de que una misma muestra de sujetos aporte más de una estimación de la fiabilidad (p. ej., consistencia interna, estabilidad temporal, formas paralelas).

El proceso de codificación de las características de los estudios y el de obtención de los coeficientes de fiabilidad son tareas sujetas a un cierto nivel de subjetividad, por lo que es muy recomendable someterlas a un estudio de fiabilidad que permita valorar si se ejecutaron con la precisión apropiada. Para ello, un procedimiento económico en tiempo y recursos consiste en seleccionar una muestra aleatoria de todos los estudios del metaanálisis y llevar a cabo un proceso de codificación doble de estas labores por parte de dos codificadores independientes.

Análisis estadístico e interpretación

Una vez que se han codificado los estudios, el paso siguiente consiste en el análisis estadístico de los datos. Los propios precursores de este enfoque no han planteado líneas concretas de análisis[39,42,43], lo que ha llevado a que exista cierta diversidad en los análisis estadísticos que se han aplicado en los estudios de GF publicados hasta la fecha.

Las distintas propuestas difieren en cuanto a[1,44–47]: a) la conveniencia de ponderar o no cada coeficiente de fiabilidad por algún factor, como el tamaño muestral o la inversa de la varianza de dicho coeficiente; b) la conveniencia de transformar el coeficiente de fiabilidad a una métrica diferente que logre asegurar el supuesto de normalidad de la distribución y estabilizar la variabilidad (p. ej., la transformación *Z* de Fisher); c) el modelo estadístico subyacente (efectos fijos, aleatorios o mixtos), y d) el modo de comprobar el influjo de variables moderadoras (p. ej., aplicando contrastes de hipótesis convencionales o no convencionales).

Pese a esta diversidad de opciones a la hora de realizar los análisis, existe un consenso en cuanto al modo de estructurarlos en función de cuatro objetivos básicos: 1) descripción de las características de los estudios; 2) estimación de la fiabilidad media; 3) evaluación de la heterogeneidad de las estimaciones de la fiabilidad, y 4) si existe heterogeneidad, búsqueda de variables moderadoras que permitan dar cuenta de tal variabilidad.

1) Descripción de las características de los estudios
   El primer objetivo de un estudio de GF es describir las características de los grupos sobre las que se ha aplicado el test, las diferentes versiones o adaptaciones del test y los diferentes contextos o propósitos para los que el test se ha aplicado. Esta descripción permite, además, ofrecer al lector una especie de fotografía de cuál es el estudio prototípico en el que se ha aplicado el test. Para alcanzar este objetivo, se utilizan técnicas estadísticas descriptivas (p. ej., medias y desviaciones típicas) y gráficas (diagramas de barras o de sectores, histogramas, gráfico en tronco y hojas [*stem-and-leaf display*] o el gráfico de caja [*boxplot*]).
2) Estimación de la fiabilidad media
   Una vez que hemos obtenido los coeficientes de fiabilidad se calcula un coeficiente de fiabilidad promedio que reflejará el nivel global medio de la fiabilidad obtenida por las aplicaciones del test. En este análisis es importante tener en cuenta que no se deben mezclar coeficientes de fiabilidad que se hayan calculado a partir de distintos métodos de estimación de la fiabilidad

(test-retest, formas paralelas o dos mitades), ya que se obtienen a partir de diferentes concepciones del error de medida[47].

También resulta problemático mezclar en un mismo análisis varios coeficientes de fiabilidad obtenidos en una misma muestra, incluso aunque sean del mismo tipo. En este caso, la decisión que debe tomarse es claramente evitar este tipo de prácticas, ya que violan el supuesto de independencia propio de las técnicas del metaanálisis.

Una característica importante de la distribución de los coeficientes de fiabilidad, independientemente del método utilizado para su obtención, es que su distribución es sesgada. Así, algunos investigadores abogan por su transformación[18,25,48,49], mientras que otros son partidarios de analizar la forma original del coeficiente de fiabilidad[19,46,50].

En nuestra opinión, los coeficientes de fiabilidad que se calculan como si fueran coeficientes de correlación de Pearson (p. ej., fiabilidad test-retest y formas paralelas) pueden transformarse a *Z* de Fisher para lograr una mejor aproximación a la distribución normal, y estabilizar las varianzas mediante la ecuación siguiente:

$$Z_i = \frac{1}{2}\log_e\left(\frac{1+r_i}{1-r_i}\right), \tag{1}$$

donde r_i es el coeficiente de fiabilidad estimado en la iésima muestra, y Z_i el coeficiente transformado. Sin embargo, si se quiere metaanalizar coeficientes alfa es más apropiado emplear la transformación de raíz cúbica derivada por Hakstian y Whalen[51]:

$$T_i = (1-r_i)^{1/3}, \tag{2}$$

donde T_i es el coeficiente transformado.

A partir de un conjunto de *k* coeficientes de fiabilidad, r_i, la estimación media de la fiabilidad, r_+, se obtendrá mediante:

$$r_+ = \frac{\sum_i w_i r_i}{\sum_i w_i}, \tag{3}$$

donde w_i es el factor de ponderación asignado a cada coeficiente de fiabilidad y r_i es el coeficiente de fiabilidad (transformado a partir de T_i, Z_i o sin transformar). Si w_i = 1, entonces obtendremos una media aritmética simple de los coeficientes de fiabilidad.

Aún pueden citarse otras opciones metodológicas, aunque con menor arraigo hasta la fecha. Por ejemplo, puede utilizarse el índice de fiabilidad en lugar del coeficiente de fiabilidad, calculándolo como la raíz cuadrada de este último. Esta estrategia se justifica sobre la base de que el índice de fiabilidad se define como el cociente entre dos varianzas (la de las puntuaciones verdaderas y la de las puntuaciones empíricas). Yendo más allá, también resultaría plausible aplicar la transformación *Z* de Fisher al índice de fiabilidad en lugar de al coeficiente de fiabilidad.

Las fórmulas propuestas anteriormente para hallar una estimación media de la fiabilidad tienen en común el hecho de que incorporan un factor de ponderación. La ponderación de los coeficientes proporciona una estimación más eficiente de la fiabilidad. Así lo demuestran varios estudios de simulación Monte Carlo[45–47]. Los factores de ponderación suelen estar relacionados con el tamaño muestral, ya que éste está directamente ligado a la precisión del coeficiente de fiabilidad. En concreto, y según los estudios de simulación, el factor de ponderación que logra la menor varianza de error se consigue calculando la inversa de la varianza de la distribución muestral del estadístico que estamos tratando. Si el coeficiente de fiabilidad se ha obtenido como una correlación de Pearson, el estimador de la varianza muestral de r_i será

$$S^2_{r_i} = \frac{(1-r_i^2)^2}{N_i - 2}. \tag{4}$$

Si hemos transformado los coeficientes de fiabilidad, las varianzas muestrales de Z_i y de T_i son, respectivamente[47]:

$$S^2_{Z_i} = \frac{1}{N_i - 3} \tag{5}$$

$$S^2_{T_i} = \frac{18J_i(N_i-1)(1-r_i)^{2/3}}{(J_i-1)(9N_i-11)^2,} \tag{6}$$

siendo J_i el número de ítems del test. Por tanto, cuando queremos ponderar cada coeficiente de fiabilidad por la inversa de su varianza muestral, hacemos que el valor de cada ponderación quede definido como:

$$w_i = \frac{1}{S^2_{r_i}} = \frac{1}{S^2_{Z_i}} = \frac{1}{S^2_{T_i}}, \tag{7}$$

según que estemos integrando coeficientes de fiabilidad, *Z* de Fisher o transformaciones *T*, respectivamente. En definitiva, la estimación de la fiabilidad media de un conjunto de *k* muestras puede adoptar distintas formas dependiendo de que deseemos o no ponderar las estimaciones, o de que queramos ponderar por el tamaño muestral o por la inversa de la varianza de cada estimación. De todas estas opciones, nuestra recomendación es utilizar la transformación *Z* de Fisher cuando el coeficiente de fiabilidad en cuestión se calcule como una correlación de Pearson, y utilizar la transformación *T* para los coeficientes de fiabilidad de consistencia interna. No recomendamos el uso directo de los coeficientes de fiabilidad porque su distribución muestral será necesariamente asimétrica[51,52]. Aunque las distribuciones *Z* de Fisher y *T* no logran normalizar por completo la distribución muestral del estadístico, se acercan bastante a ella y, en consecuencia, son soluciones preferibles[47]. Por último, junto con la estimación de la fiabilidad media se suele calcular un *intervalo de confianza* asumiendo una distribución normal.

El procedimiento de cálculo del coeficiente de fiabilidad medio asume un modelo de efectos fijos, según el cual todos los coeficientes de fiabilidad están estimando a un único coeficiente de fiabilidad paramétrico, común a todos ellos, de forma que la variabilidad observada entre ellos se debe exclusivamente al error de muestreo aleatorio[53–55]. Otro modelo estadístico aplicable en

metaanálisis es el modelo de efectos aleatorios, que implica asumir que los coeficientes de fiabilidad obtenidos en los estudios estiman a una distribución de coeficientes de fiabilidad paramétricos, de forma que la variabilidad observada entre ellos es la suma de la variabilidad provocada por el error de muestreo y la varianza intercoeficientes, τ^2. Si el modelo asumido es el de efectos aleatorios[56,57], entonces el cálculo del coeficiente de fiabilidad promedio se obtiene utilizando como factor de ponderación la inversa de la varianza, que en este caso es la suma de dos varianzas: la varianza intraestudio y alguna estimación de la varianza intercoeficientes[58–60]. La decisión de asumir un modelo u otro debe hacerse sobre una base conceptual y también puede ayudar a dicha decisión el análisis de la heterogeneidad presentada a continuación.

3) Evaluación de la heterogeneidad
   Después de obtener una estimación media de la fiabilidad, el siguiente objetivo de un estudio de GF consiste en valorar el grado de heterogeneidad existente entre los coeficientes individualmente reportados por estudios distintos. Esta es una fase fundamental de cara a las conclusiones del estudio. Si después de este análisis los coeficientes resultan homogéneos, podremos concluir entonces que la estimación de la fiabilidad media obtenida anteriormente es generalizable a cualquier aplicación del test que estemos estudiando. Esta conclusión es la que prematuramente se adopta cuando se aplica el test sin calcular una estimación de la fiabilidad de las puntuaciones. Por lo general, sin embargo, esta suposición idealista se vuelve insostenible cuando interpretamos los resultados de las pruebas de homogeneidad, siempre que el número de estudios otorgue a nuestro análisis una potencia estadística apropiada.
   El procedimiento más apropiado para determinar si un conjunto de coeficientes de fiabilidad es homogéneo consiste en aplicar el estadístico Q de heterogeneidad, que se obtiene mediante:

$$Q = \sum_i w_i (r_i - r_+)^2, \tag{8}$$

   teniendo en cuenta que en dicha ecuación r_i y r_+ pueden sustituirse por Z_i y Z_+, o bien por T_i y T_+, según el índice que se esté utilizando en el metaanálisis. El factor de ponderación, w_i, viene definido por la ecuación 7. Un resultado significativo para el estadístico Q implicará asumir que los coeficientes de fiabilidad varían entre sí más de lo que el error de muestreo aleatorio es capaz de explicar. No obstante, dado que el estadístico Q tiene baja potencia estadística con un número reducido de coeficientes ($k<30$)[61,62], se recomienda complementarlo con el índice I^2, un estadístico que describe en tantos por ciento qué parte de la variabilidad observada entre los coeficientes de fiabilidad se debe a verdadera heterogeneidad provocada por factores que van más allá del mero error de muestreo[63,64]. El índice I^2 se obtiene mediante la ecuación[55,56]:

$$I^2 = \frac{Q - (k-1)}{Q} \times 100. \tag{9}$$

4) Búsqueda de variables moderadoras
   Si existe heterogeneidad entre los coeficientes de fiabilidad, se hace preciso buscar variables moderadoras que den cuenta de dicha variabilidad. Tomando los moderadores como variables independientes (o predictoras) y los coeficientes de fiabilidad (o su transformación a Z o T) como variable dependiente, se pueden aplicar contrastes de hipótesis, tales como ANOVA cuando la variable independiente es cualitativa (p. ej., el idioma en que se aplicó el test), y análisis de regresión cuando es continua (p. ej., la desviación típica de las puntuaciones del test). Pero en lo que no existe consenso hasta ahora es en el modelo estadístico desde el que aplicar tales contrastes de hipótesis. Así, los primeros estudios de GF aplicaron las *técnicas convencionales* de ANOVA y de regresión, es decir, sin ponderar los coeficientes de fiabilidad en función de la precisión (es decir, haciendo $w_i = 1$). Sin embargo, posteriormente se han aplicado procedimientos de ponderación asumiendo *modelos de efectos fijos con moderadores*. Pero actualmente, se consideran más apropiados los *modelos de efectos mixtos*, según los cuales el factor de ponderación debe incorporar tanto una estimación de la varianza muestral del coeficiente como de la varianza intercoeficientes, τ^2, actuando la variable moderadora como un factor de efectos fijos.

Publicación

Una vez concluida la fase de análisis de datos e interpretación de los resultados, sólo nos queda emprender la redacción formal del estudio para su posterior publicación.

La estructura que emplearemos será similar a la de cualquier investigación, a saber: introducción, método, resultados y discusión. En lo referente a los subapartados que deben tratarse en cada uno de estos epígrafes, el formato que seguiremos será el que habitualmente se utiliza en la publicación de revisiones metaanalíticas[15,65,66].

En la introducción debe hablarse, en primer lugar, del test objeto de análisis, así como de sus posibles versiones y campos de aplicación. Este apartado debe dejar claro al lector que existen motivos para estudiar el instrumento en cuestión, lo cual suele respaldarse por el hecho de que los test escogidos son de uso muy extendido, con lo que la masa social a la que pueden resultar de interés las conclusiones del trabajo es muy amplia. En la sección del método deben detallarse con la mayor minuciosidad posible las decisiones tomadas a lo largo del estudio, lo cual garantizará una máxima transparencia y la posibilidad de que otros investigadores puedan replicar el trabajo. Lo habitual es presentar las secciones típicas de un estudio metaanalítico: a) definición de los criterios de selección de los estudios empíricos para el metaanálisis; b) descripción de los procedimientos de búsqueda de los estudios (bases de datos electrónicas consultadas, palabras clave utilizadas, otras estrategias de búsqueda y resultado del proceso de búsqueda); c) identificación de las características (metodológicas, de contexto, sustantivas y extrínsecas) de los estudios que se van a registrar para comprobar su posible relación con los coeficientes de fiabilidad; d) descripción de los diferentes coeficientes de fiabilidad que se registraron, y

e) especificación de las técnicas de análisis estadístico utilizadas.

La sección de resultados debe comenzar con un apartado descriptivo donde se detallen las características de los estudios incluidos en el metaanálisis. Después se presentará una estimación de la fiabilidad media de la escala —y, en su caso, también de las subescalas— separando, si se hubiesen recogido coeficientes de distinta naturaleza, los promedios para cada uno de ellos. Seguidamente se constatará si existe heterogeneidad entre las estimaciones de la fiabilidad para, si la respuesta fuese afirmativa (como de hecho suele ocurrir), proceder a los análisis estadísticos utilizando como predictores las variables moderadoras previamente codificadas. La variable criterio será, por lo general, la estimación de la fiabilidad de las puntuaciones del test de cada estudio. Es aconsejable acompañar esta sección de tablas y gráficos. En la sección de discusión y de conclusiones se deben relacionar los resultados obtenidos con los de otros estudios de GF similares, así como ofrecer una valoración de la fiabilidad promedio que ofrecen las puntuaciones del test, la heterogeneidad encontrada entre los coeficientes y las variables moderadoras que se han mostrado relacionadas con esta variabilidad.

Finalmente, en la sección de referencias deben destacarse de algún modo (por ejemplo, con un asterisco) los artículos empleados en el metaanálisis. Además, y siempre que el espacio lo permita, resultaría útil incluir un apéndice con la base de datos completa, en la que aparezcan las principales variables que han sido utilizadas en la fase de análisis estadístico.

Conclusión

El propósito de este artículo fue presentar una panorámica de qué es el enfoque metaanalítico de GF, definido como una reciente metodología que tiene por objeto integrar cuantitativamente las estimaciones de la fiabilidad obtenidas en aplicaciones sucesivas de un determinado test o conjunto de instrumentos de medida con objeto de determinar en qué medida dichas estimaciones varían de una muestra a otra y cuáles pueden ser los factores y características de los estudios y de las muestras que explican tal variabilidad. Hemos presentado cuáles son las etapas mediante las que se lleva a cabo un estudio de esta naturaleza y cuáles son los aspectos estadísticos y psicométricos de este enfoque que actualmente son objeto de estudio y discusión.

En la raíz de este enfoque metodológico se encuentra la crítica, planteada en los últimos años por numerosos autores, contra la idea errónea y muy extendida entre los investigadores y los profesionales en ciencias de la salud de que la fiabilidad es una propiedad del test, cuando realmente es una propiedad inherente a las puntuaciones obtenidas en una determinada aplicación del test. Frases del tipo ''la fiabilidad del test es de 0,80'', son incorrectas. Lo correcto es decir ''la fiabilidad de las puntuaciones del test sobre esta muestra es de 0,80''.

En consecuencia, los investigadores en ciencias de la salud y otros campos afines debemos ser cada vez más conscientes de la necesidad de estimar la fiabilidad alcanzada por las puntuaciones del test en la propia muestra y no inducirla a partir de aplicaciones previas del test. Aunque esta metodología se encuentra todavía en fase de depuración, es indiscutible el importante papel que están jugando los estudios de esta naturaleza para concienciar a la comunidad científica de la importancia de considerar la fiabilidad como una cuestión empírica que tiene que estimarse con los datos de las propias muestras y evitar inducciones que pueden provocar serios errores en la estimación de la precisión de nuestras medidas.

Financiación

Este artículo ha sido financiado por el Fondo de Investigación Sanitaria, convocatoria de Evaluación de Tecnologías Sanitarias (Proyecto Nº: PI07/90384).

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Bibliografía

1. Onwuegbuzie AJ, Daniel LG. Reliability generalization: The importance of considering sample specificity, confidence intervals, and subgroup differences. Research in the Schools. 2004b; 11:60–71.
2. Crocker L, Algina J. Introduction to classical and modern test theory. Nueva York: Holt, Rinehart, & Winston; 1986.
3. Gronlund NE, Linn RL. Measurement and evaluation in teaching (6 ed.). New York: Macmillan; 1990.
4. Pedhazur EJ, Schmelkin LP. Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum; 1991.
5. Rowley GL. The reliability of observational measures. Am Educ Res J. 1976b;13:51–9.
6. Henson RK, Thompson B. Characterizing measurement error in scores across studies: Some recommendations for conducting ''reliability generalization'' studies. Measurement and Evaluation in Counseling and Development. 2002b;35:113–27.
7. Vacha-Haase T, Kogan LR, Thompson B. Sample compositions and variabilities in published studies versus those in test manuals. Educ Psychol Meas. 2000b;60:509–22.
8. Vacha-Haase T, Ness C. Practices regarding reporting of reliability coefficients: A review of three journals. J Exp Educ. 1999b;67:335–42.
9. Whittington D. How well do researchers report their measures? An evaluation of measurement in published educational research Educ Psychol Meas. 1998b;58:21–37.
10. Vacha-Haase T, Henson RK, Caruso JC. Reliability generalization: Moving toward improved understanding and use of score reliability. Educ Psychol Meas. 2002b;62:562–9.
11. Wilkinson L, APA Task Force on Statistical Inference. Statistical methods in psychology journal: Guidelines and explanations. Am Psychol. 1999b;54:594–604.
12. Nunnally JC. Reliability of measurement. In: Mitzel HE, editor. Encyclopedia of educational research. New York: Free Press; 1982. p. 1589–601.
13. Thompson B. Guidelines for authors. Educ Psychol Meas. 1994b; 54:837–47.
14. Heldref Foundation. Guidelines for contributors. J Exp Educ. 1997b;65:95–6.
15. Botella J, Gambara H. ¿Qué es el meta-análisis? Madrid: Biblioteca Nueva; 2002.
16. Cooper HM. Integrating research: A guide for literature reviews, 2 ed. Thousand Oaks, CA: Sage; 1998.

17. Cooper H, Hedges LV, editors. The handbook of research synthesis. New York: Russell Sage Foundation; 1994.
18. Glass GV, McGaw B, Smith ML. Meta-analysis in social research. Beverly Hills, CA: Sage; 1981.
19. Hedges LV, Olkin I. Statistical methods for meta-analysis. Orlando, FL: Academic Press; 1985.
20. Hunter JE, Schmidt FS. Methods of meta-analysis: Correcting error and bias in research findings, 2 ed. Thousand Oaks, CA: Sage; 2004.
21. Martín JLR, Tobías A, Seoane T, coordinadores. Revisiones sistemáticas en las ciencias de la vida. Toledo: FISCAM; 2006.
22. Petticrew M, Roberts H. Systematic reviews in the social sciences: A practical guide. Malden, MA: Blackwell; 2006.
23. Sánchez-Meca J, Ato M. Meta-análisis: una alternativa metodológica a las revisiones tradicionales de la investigación. In: Arnau J, Carpintero H, editors. Tratado de psicología general I: historia, teoría y método. Madrid: Alhambra; 1989. p. 617–69.
24. Schulze R. Meta-analysis: A comparison of approaches. Göttingen: Hogrefe & Huber Pub; 2004.
25. Dimitrov DM. Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. Educ Psychol Meas. 2002b;62:783–801.
26. Sawilowsky SS. Psychometrics versus datametrics: Comment on Vacha-Haase's 'Reliability generalization' method and some *EPM* editorial policies. Educ Psychol Meas. 2000b;60:157–173.
27. Sawilowsky SS. Reliability: Rejoinder to Thompson and Vacha-Haase. Educ Psychol Meas. 2000b;60:196–200.
28. Yin P, Fan X. Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. Educ Psychol Meas. 2000b;60:201–23.
29. Barnes LLB, Harp D, Jung WS. Reliability generalization of scores on the Spielberger State-Trait Anxiety Inventory. Educ Psychol Meas. 2002b;62:603–18.
30. Campbell JS, Pulos S, Hogan M, Murry F. Reliability generalization of the Psychopathy Checklist applied in youthful samples. Educ Psychol Meas. 2005b;65:639–56.
31. Li A, Bagger J. The Balanced Inventory of Desirable Responding (BIDR): A reliability generalization study. Educ Psychol Meas. 2007b;67:525–44.
32. Beretvas SN, Suizzo M-A, Durham JA, Yarnell LM. A reliability generalization study of scores on Rotter's and Nowicki-Strickland's locus of control scales. Educ Psychol Mea. 2008b;68:97–119.
33. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951b;15:297–334.
34. Lipsey MW, Wilson DB. Practical meta-analysis. Thousand Oaks, CA: Sage; 2001.
35. Marín-Martínez F, Sánchez-Meca J, Huedo T, Fernández I. Meta-análisis: ¿Dónde estamos y hacia dónde vamos? In: Borges A, Prieto P, editors. Psicología y ciencias afines en los albores del siglo XXI (Homenaje al profesor Alfonso Sánchez Bruno). Tenerife: Grupo Editorial Universitario; 2007.
36. Rosenthal R. Meta-analytic procedures for social research, 2 ed. Newbury Park, CA: Sage; 1991.
37. Sánchez-Meca J. La revisión del estado de la cuestión: el meta-análisis. En: Camisón C, Oltra MJ, Flor ML, editores. Enfoques, problemas y métodos de investigación en economía y dirección de empresas. Castellón: ACEDE/Fundació Universitat Jaime I–Empresa; 2003. p. 101–110.
38. Leach LF, Henson RK, Odom LR, Cagle LS. A reliability generalization study of the Self-Description Questionnaire. Educ Psychol Meas. 2006b;66:285–304.
39. Petrosino A, Boruch RF, Soydan H, Duggan L, Sánchez-Meca J. Meeting the challenges of evidence-based policy: The Campbell Collaboration. Ann Am Acad Pol Soc Sci. 2001b;578:14–34.
40. Sánchez-Meca J, Boruch RF, Petrosino A, Rosa-Alcázar AI. La Colaboración Campbell y la práctica basada en la evidencia. Papeles del Psicólogo. 2002b;83:44–8.
41. Shadish WR, Chacón-Moscoso S, Sánchez-Meca J. Evidence-based decision making: Enhancing systematic reviews of program evaluation results in Europe. Evaluation. 2005b;11:95–109.
42. Thompson B, editor. Score reliability: Contemporary thinking on reliability issues. Thousand Oaks, CA: Sage; 2003.
43. Vacha-Haase T. Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educ Psychol Meas. 1998b;58:6–20.
44. Beretvas SN, Pastor DA. Using mixed-effects models in reliability generalization studies. Educ Psychol Meas. 2003b;63:75–95.
45. Feldt LS, Charter RA. Averaging internal consistency reliability coefficients. Educ Psychol Meas. 2006b;66:215–27.
46. Mason C, Allam R, Brannick MT. How to meta-analyze coefficient-of-stability estimates: Some recommendations based on Monte Carlo studies. Educ Psychol Meas. 2007b;67:765–83.
47. Rodríguez MC, Maeda Y. Meta-analysis of coefficient alpha. Psychol Methods. 2006b;11:306–22.
48. Silver N, Dunlap W. Averaging coefficients: Should Fisher's z-transformation be used? J Appl Psychol. 1987b;72:3–9.
49. Thompson B, Vacha-Haase T. Psychometrics *is* datametrics: The test is not reliable. Educ Psychol Meas. 2000b;60:174–95.
50. Hall SM, Brannick MT. Comparison of two random-effects methods of meta-analysis. J Appl Psychol. 2002b;87:377–89.
51. Hakstian AR, Whalen TE. A *k*-sample significance test for independent alpha coefficients. Psychometrika. 1976b;41: 219–231.
52. Feldt LS, Brennan RL. Reliability. In: Linn RL, editor. Educational measurement, 3 ed. New York: American Council on Education and Macmillan; 1989. p. 105–46.
53. Hedges LV. Fixed effects models. In: Cooper H, Hedges LV, editors. The handbook of research synthesis. New York: Russell Sage Foundation; 1994. p. 285–99.
54. Marín-Martínez F, Sánchez-Meca J. Testing dichotomous moderators in meta-analysis. J Exp Educ. 1998b;67:69–81.
55. Sánchez-Meca J, Marín-Martínez F. Testing continuous moderators in meta-analysis: A comparison of procedures. Br J Math Stat Psychol. 1998b;51:311–26.
56. Hedges LV, Vevea JL. Fixed- and random-effects models in meta-analysis. Psychol Methods. 1998b;3:486–504.
57. Sánchez-Meca J, Marín-Martínez F, Huedo-Medina T. Modelo de efectos fijos y modelo de efectos aleatorios. En: Martín JLR, Tobías A, Seoane T. (Coords.), Revisiones sistemáticas en ciencias de la vida. Toledo: FISCAM; 2006. p. 189–204.
58. Sánchez-Meca J, Marín-Martínez F. Confidence intervals for the overall effect size in random-effects meta-analysis. Psychol Methods. 2008b;13:31–48.
59. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. J Educ Behav Stat. 2005b;30:261–93.
60. Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. Statistics in Medicine. 2007b;26:37–52.
61. Harwell M. An empirical study of Hedges's homogeneity test. Psychol Methods. 1997b;2:219–31.
62. Sánchez-Meca J, Marín-Martínez F. Homogeneity tests in meta-analysis: A Monte Carlo comparison of statistical power and Type I error. Quality and Quantity. 1997b;31: 385–399.
63. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine. 2002b;21:1539–58.
64. Huedo-Medina T, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: *Q* statistic or I^2 index? Psychol Methods. 2006b;11:193–206.