Escritos de Psicología
2008, 2-1: 110-121

UNA REVISIÓN DE LOS ESTUDIOS META-ANALÍTICOS DE GENERALIZACIÓN DE LA FIABILIDAD

A REVIEW OF THE META-ANALYTIC STUDIES OF RELIABILITY GENERALIZATION

Julio Sánchez-Meca, José A. López-Pina
y José A. López-López
Dpto. Psicología Básica y Metodología. Facultad de Psicología.
Universidad de Murcia
e-mail: *jsmeca@um.es*

Resumen El enfoque meta-analítico de generalización de la fiabilidad (GF) pretende demostrar que la fiabilidad es una propiedad empírica que varía de una aplicación a otra del test. Este nuevo enfoque meta-analítico está contribuyendo a concienciar a los investigadores sobre la importancia de aportar estimaciones de la fiabilidad con los propios datos y evitar inducciones de la fiabilidad. Se presentan las fases en las que se lleva a cabo un estudio GF: (a) formulación del problema, (b) búsqueda de los estudios, (c) codificación de los estudios y (d) análisis estadístico e interpretación. Se presenta una visión actualizada de los problemas estadísticos de este enfoque: (a) transformar versus no transformar los coeficientes de fiabilidad, (b) ponderar versus no ponderar los coeficientes y (d) cuál es el modelo estadístico más apropiado (efectos fijos, efectos aleatorios, efectos mixtos). Se presenta una revisión sistemática de los 49 estudios GF publicados hasta la fecha y se analiza la variabilidad en el modo de analizar estadísticamente los datos. Finalmente, se discuten las implicaciones de los estudios GF para la investigación y la práctica profesional.

Palabras clave generalización de la fiabilidad, meta-análisis, coeficiente de fiabilidad.

Abstract The meta-analytic approach of reliability generalization (RG) pretends to show that reliability is an empirical property that varies from one test application to another. This recent meta-analytic approach is helping to make the researchers aware of the importance of reporting reliability estimates obtained from the own data and, of avoiding the malpractice of inducting reliability coefficients from other studies and previous applications of the test. The stages to carry out an RG study are presented: (a) formulating the problem, (b) searching for the studies, (c) coding studies, and (d) statistical analysis and interpretation. An updated overview of the statistical problems of this approach is also offered: (a) to transform versus not to transform the reliability coefficients, (b) to weight versus not to weight the coefficients, and (d) which statistical model is the most appropriate (fixed-, random-, and mixed-effects). A systematic review of the 49 RG studies published to date is presented with the purpose of analyzing the heterogeneity in how the data are statistically analyzed. Finally, the implications of the RG studies for research and professional practice are discussed.

Key words Reliability generalization; meta-analysis; reliability coefficient.

Una de las herramientas fundamentales en la práctica profesional y en la investigación psicológica es el uso de tests e instrumentos de medida para cuantificar el nivel exhibido por las personas en los constructos psicológicos. Conocer la calidad métrica de los instrumentos de medida, tales como la fiabilidad y la validez, constituye una tarea esencial tanto para el profesional de la psicología como para el investigador en las ciencias del comportamiento.

Una de las propiedades psicométricas que debe poseer un instrumento de medida psicológica es la fiabilidad. Los tests psicológicos debidamente baremados ofrecen estimaciones de su calidad métrica, de forma que es muy común que los profesionales y los investigadores asuman que la fiabilidad de las puntuaciones en la muestra que ellos han analizado sea la misma que la obtenida en el proceso de baremación original del test. Ello hace que resulte muy frecuente encontrar frases del tipo "el coeficiente de fiabilidad del test es 0.80". Sin embargo, la fiabilidad no es una propiedad del test, sino de las puntuaciones obtenidas cada vez que se aplica el test a una muestra (Crocker y Algina, 1986; Gronlund y Linn, 1990; Traub, 1994). De hecho, la fiabilidad puede variar ostensiblemente de una aplicación a otra dependiendo de diversos factores, tales como la variabilidad de las puntuaciones en la muestra, la población de procedencia de las personas que la componen, el propósito de la aplicación del test o el contexto donde se aplica.

Esta práctica, generalizada entre los investigadores y clínicos, de asumir para su propia muestra la fiabilidad obtenida en alguna aplicación previa del test ha sido denominada por Vacha-Haase, Kogan y Thompson (2000) como *inducción de la fiabilidad* (reliability induction), donde el término 'inducción' hace referencia al hecho de que el investigador atribuye a las puntuaciones de su muestra la fiabilidad obtenida en alguna aplicación previa del test a otra muestra de sujetos con características presumiblemente similares, como si la fiabilidad obtenida en dicha aplicación del test fuera extrapolable a otras aplicaciones del mismo. Diversos estudios han demostrado lo arraigada que está esta práctica en la comunidad científica. Así, Vacha-Haase y Ness (1999), al revisar los estudios empíricos publicados en las revistas *Journal of Counseling Psychology*, *Psychology and Aging* y *Professional Psychology: Research and Practice*, encontraron que sólo el 35.6% de los artículos aportaron alguna estimación de la fiabilidad a partir de los propios datos analizados, mientras que el resto inducía la fiabilidad a partir de aplicaciones previas del test, o bien no hacían alusión alguna a la fiabilidad. Así mismo, Whittington (1998) revisó los estudios empíricos publicados en 22 revistas del ámbito de la educación y encontró que el 54% de éstos indujeron la fiabilidad desde otras aplicaciones de los tests.

Contra esta extendida práctica de inducir la fiabilidad se está promoviendo desde diversos foros científicos la necesidad de que los investigadores reporten estimaciones originales de la fiabilidad de los instrumentos de medida utilizados con los propios datos de la muestra. Prueba de ello son las recomendaciones que en este sentido ha propuesto la *APA Task Force on Statistical Inference* (Wilkinson y APA Task Force on Statistical Inference, 1999), asociaciones científicas tales como la *American Educational Research Association* y el *National Research Council on Measurement in Education*, así como desde las políticas editoriales de algunas revistas, tales como *Educational and Psychological Measurement* (Thompson, 1994) o *Journal of Experimental Education* (Heldref Foundation, 1997).

En lugar de asumir que la fiabilidad de un test es homogénea o, lo que es lo mismo, generalizable a través de las diferentes aplicaciones del mismo a diferentes muestras de personas, la cuestión de la generalización de la fiabilidad es una cuestión empírica que puede abordarse mediante la integración cuantitativa de los coeficientes de fiabilidad que las aplicaciones de un determinado test ha dado lugar en estudios empíricos realizados por diferentes investigadores, con diferentes muestras y en diferentes contextos de aplicación. No hace mucho que Vacha-Haase (1998) ha propuesto una metodología capaz de llevar a cabo esta tarea de integración cuantitativa, que se ha denominado el enfoque meta-analítico de *generalización de la fiabilidad* (reliability generalization). Gracias a este método de revisión es posible comprobar si la fiabilidad exhibida por las aplicaciones de un test en múltiples muestras es generalizable o si, por el contrario, las estimaciones de la fiabilidad se muestran heterogéneas entre si. Sin embargo, y principalmente debido a su corta historia, no existe un modo monolítico de cómo analizar estadísticamente la integración de las estimaciones de la fiabilidad, lo que ha provocado que los estudios de generalización de la fiabilidad (GF) muestren claras divergencias en los métodos estadísticos aplicados.

El propósito de este artículo es presentar una panorámica de cómo se lleva a cabo un estudio GF y, dada la heterogeneidad de métodos de integración estadística que se han aplicado en este tipo de estudios, presentamos una revisión sistemática de las características metodológicas de los 49 estudios GF que hasta la fecha se han publicado.

EL ENFOQUE DE GENERALIZACIÓN DE LA FIABILIDAD

La integración cuantitativa preconizada por Vacha-Haase (1998) de coeficientes de fiabilidad obtenidos en las aplicaciones de un test a diferentes muestras es un tipo de meta-análisis en el que, en lugar de integrar estimacio-

nes del tamaño del efecto (e.g., diferencias de medias estandarizadas, odds ratios, etc.), se integran las estimaciones de la fiabilidad obtenidas con los propios datos de las muestras de personas. El análisis estadístico de los coeficientes de fiabilidad integrados permite: (a) ofrecer una estimación de la fiabilidad media de las aplicaciones del test, (b) comprobar si los coeficientes de fiabilidad son homogéneos entre si, es decir, si las discrepancias entre ellos se pueden deber a mero error de muestreo aleatorio o si, por el contrario, los coeficientes muestran una heterogeneidad que el error de muestreo por si solo no puede explicar y (c) caso de que las estimaciones de la fiabilidad sean heterogéneas, se examina el influjo de variables moderadoras de tal heterogeneidad (Henson y Thompson, 2002; Onwuegbuzie y Daniel, 2004; Rodriguez y Maeda, 2006; Thompson, 2003; Vacha-Haase, 1998; Vacha-Haase, Henson y Caruso, 2002).

Los estudios GF pueden ofrecer una importante contribución hacia una mejor comprensión de los factores que influyen en la fiabilidad de las puntuaciones de los tests y a concienciar a los investigadores y profesionales de la psicología sobre la necesidad de hacer estimaciones de la fiabilidad de los instrumentos de medida con los propios datos de la muestra y evitar la arriesgada práctica de la inducción de la fiabilidad.

Desde su inicio en 1998 hasta la fecha, ya se han publicado 49 estudios GF sobre muy diversos tests psicológicos e instrumentos de medida. Como ejemplos, cabe mencionar los estudios GF realizados sobre el *Beck Depression Inventory* (Yin y Fan, 2000), el *Spielberger State-Trait Anxiety Inventory* (Barnes, Harp y Jung, 2002), el *Psychopathy Checklist* (Campbell, Pulos, Hogan y Murry, 2005), el *Balanced Inventory of Desirable Responding* (Li y Bagger, 2007) o las escalas de locus de control de Rotter y de Nowicki-Strickland (Beretvas, Suizzo, Durham y Yarnell, 2008).

Un estudio GF es un tipo de meta-análisis y, como tal, se lleva a cabo mediante cuatro etapas claramente delimitadas: (a) formulación del problema, (b) búsqueda de los estudios, (c) codificación de los estudios y (d) análisis estadístico e interpretación (Cooper y Hedges, 1994; Hunter y Schmidt, 2004; Sánchez-Meca, 2003; Sánchez-Meca y Ato, 1989).

1. Formulación del problema

El propósito de un estudio GF es examinar cómo varían las estimaciones de la fiabilidad obtenidas cuando un determinado test se ha aplicado en múltiples estudios con muestras de sujetos diferentes que pueden proceder de poblaciones diferentes y encontrar los factores o características de los estudios que pueden dar cuenta de la variabilidad exhibida por dichas estimaciones de la fiabilidad. Las estimaciones de la fiabilidad pueden variar en función de la variabilidad de las puntuaciones del test en cada muestra. Además, pueden existir adaptaciones del test a otros idiomas, culturas o países, y también pueden existir varias versiones del test (e.g., una versión abreviada con un número reducido de ítems) o puede haberse adaptado a diferentes edades de los sujetos. Todos estos factores, y otros muchos más, pueden afectar a la heterogeneidad de los coeficientes de fiabilidad y su influjo puede ser el objeto de un estudio de generalización de la fiabilidad. No obstante, para que tenga sentido realizar un estudio GF es preciso que el test en cuestión tenga una amplia aplicación y que exista un número razonable de estudios empíricos que han realizado estimaciones propias de la fiabilidad (Henson y Thompson, 2002).

2. Búsqueda de los estudios

Para iniciar la búsqueda de los estudios es preciso previamente definir los criterios de selección que tienen que cumplir los estudios para ser incluidos en el estudio GF. Aunque estos criterios dependerán del propósito del meta-análisis, algunos criterios de selección son de obligada consideración. Por ejemplo, los años inicial y final del proceso de búsqueda y si se incluirán adaptaciones del test a otros idiomas, culturas, edades, longitudes, etc., o si nuestro estudio se centrará sólo en aplicaciones del test original.

La búsqueda de los estudios que cumplan con los criterios de selección deberá ser lo más completa posible y ello pasa necesariamente por el uso de las nuevas tecnologías de la información (consulta de índices de citación electrónicos) combinadas con la consulta de revistas especializadas y de expertos en el campo para localizar tanto estudios publicados como no publicados que puedan haber utilizado el test. El conjunto final de trabajos incluidos en nuestro estudio 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.

3. Codificación de los estudios

Con objeto de comprobar qué características de los estudios pueden afectar a la variabilidad de los coeficientes de fiabilidad que han aplicado el test, es preciso identificar a priori dichas características y recogerlas en un protocolo de codificación que se aplica a cada estudio empírico que aporte alguna estimación propia de la fiabilidad de las puntuaciones del test. Tales característi-

cas se pueden clasificar en metodológicas, o propias de la metodología del estudio y de las propiedades psicométricas de los tests, y sustantivas, o relativas al campo de investigación en el que suele aplicarse el test. De entre las características metodológicas cabe mencionar las diferentes formas de aplicación del test (auto-informe vs. aplicación por un evaluador), diferentes formatos de recogida de las respuestas (respuestas en papel y lápiz vs. informatizadas), diferentes 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 de la muestra y la variabilidad de las puntuaciones del test en la muestra. De entre las características sustantivas cabe mencionar la naturaleza clínica versus 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.

4. Análisis estadístico e interpretación

Una vez que disponemos de los estudios codificados en función de las características seleccionadas y que hemos registrado las estimaciones de la fiabilidad de esos estudios, se analizan estadísticamente los datos. Es en esta etapa donde los estudios GF muestran una mayor diversidad en cuanto a las técnicas estadísticas a utilizar, de forma que existe actualmente cierta controversia sobre dos cuestiones analíticas (Beretvas y Pastor, 2003; Feldt y Charter, 2006; Mason, Allam y Brannick, 2007; Rodríguez y Maeda, 2006; Sánchez-Meca y López-Pina, 2008; Sánchez-Meca, López-Pina y López-López, en prensa; Thompson, 2003; Vacha-Haase, 1998): (a) dada la típica asimetría exhibida por la distribución de los coeficientes de fiabilidad, ¿conviene transformar los coeficientes para normalizar o, al menos, simetrizar su distribución?; (b) dado que los estudios difieren en cuanto al tamaño muestral, ¿es preferible ponderar los coeficientes de fiabilidad en función del tamaño muestral o por algún otro factor de ponderación que tenga en cuenta la variabilidad de la distribución muestral de los coeficientes de fiabilidad? Y dentro de los diferentes modos de ponderación, ¿es preferible aplicar un modelo de efectos fijos o de efectos aleatorios?

Respecto de la primera cuestión, algunos autores recomiendan no transformar los coeficientes de fiabilidad, r_i, para el análisis estadístico (Hall y Brannick, 2002; Hunter y Schmidt, 2004; Onwuegbuzie y Daniel, 2004), o bien utilizar el índice de fiabilidad en lugar del coeficiente de fiabilidad ($\sqrt{r_i}$), mientras que otros proponen diferentes transformaciones (Sawilowski, 2000a; Silver y Dunlap, 1987; Thompson y Vacha-Haase, 2000), de las que la más recomendada es la propuesta por Hakstian y Whalen (1976) para integrar coeficientes de consistencia interna alfa de Cronbach, según la cual el coeficiente de fiabilidad, r_i, se transforma mediante $T_i = (1 - r_i)^{1/3}$. Otra transformación se basa en la raíz cuadrada del inverso del coeficiente de fiabilidad: $M_i = (1 - r_i)^{1/2}$. Y una última propuesta consiste en aplicar la transformación a Z de Fisher: $Z_i = \frac{1}{2}\log_e\left(\frac{1+r_i}{1-r_i}\right)$. Esta última no es aconsejable cuando los coeficientes de fiabilidad están basados en la consistencia interna del test, es decir, son coeficientes alfa de Cronbach o alguna adaptación de éste (e.g., coeficientes KR20 y KR21), ya que éstos no son estrictamente coeficientes de correlación de Pearson. Todas estas transformaciones son válidas para aquellos tipos de fiabilidad que se calculen como un coeficiente de correlación de Pearson, es decir, para las estimaciones de la fiabilidad test-retest, formas paralelas y dos mitades, aunque en este último caso no está claro si sería apropiado aplicar estas transformaciones cuando se ha calculado la corrección de Spearman-Brown.

Una vez decidido si se van a transformar los coeficientes de fiabilidad o no, el análisis estadístico se dirige a ofrecer tres respuestas: (a) estimar la fiabilidad media de las aplicaciones del test; (b) comprobar si los coeficientes de fiabilidad son más heterogéneos entre si de lo que el error de muestreo es capaz de explicar y (c) caso de ser así, explorar qué características de los estudios pueden estar afectando a la heterogeneidad de los coeficientes.

(1) *Estimación de la fiabilidad media.*

La estimación de la fiabilidad media refleja 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 provienen de los distintos métodos de estimación de la fiabilidad (test-retest, formas paralelas o consistencia interna), ya que se obtienen a partir de diferentes concepciones del error medida. La fiabilidad media se obtiene bien mediante una simple media aritmética de los coeficientes (transformados o no), o bien ponderando cada estimación de la fiabilidad en función de, por ejemplo, el tamaño muestral o la inversa de la varianza de su distribución muestral. Es decir:

$$r_+ = \frac{\sum_i w_i r_i}{\sum_i w_i} \quad , (1)$$

donde r_i puede ser sustituido por cualquiera de las transformaciones antes mencionadas (T_i, M_i, Z_i); w_i es

el factor de ponderación asignado a cada coeficiente de fiabilidad y r_i es el coeficiente de fiabilidad (transformado o no). Si $w_i = 1$, entonces obtendremos una media aritmética simple de los coeficientes de fiabilidad. Si se decide ponderar por el tamaño muestral de cada estudio, N_i, entonces $w_i = N_i$. Y si se decide ponderar por la inversa de la varianza de cada coeficiente, S_i^2, entonces $w_i = 1/S_i^2$, teniendo en cuenta que la varianza será diferente según que promediemos los coeficientes de fiabilidad, su transformación a Z de Fisher o su transformación mediante la raíz cúbica. Estas tres varianzas pueden estimarse, respectivamente, mediante:

$$S_{r_i}^2 = \frac{\left(1 - r_i^2\right)^2}{N_i - 2} \quad , (2)$$

$$S_{Z_i}^2 = \frac{1}{N_i - 3} \quad , (3)$$

$$S_{T_i}^2 = \frac{18 J_i (N_i - 1)(1 - r_i)^{2/3}}{(J_i - 1)(9N_i - 11)^2} \quad , (4)$$

siendo J_i el número de ítems del test.

Aunque las distribuciones Z de Fisher y T no logran normalizar por completo la distribución muestral del coeficiente de fiabilidad, se acercan bastante a ella y, en consecuencia, son soluciones preferibles a no transformar (Feldt y Brennan, 1989; Hakstian y Whalen, 1976; Rodríguez y Maeda, 2006). Por último, junto con la estimación de la fiabilidad media se suele calcular un *intervalo de confianza* asumiendo una distribución normal (Sánchez-Meca y Marín-Martínez, 2008).

(2) *Análisis de la heterogeneidad.*

Si los coeficientes de fiabilidad obtenidos en las aplicaciones del test son muy heterogéneos entre si (más de lo que el error de muestreo puede explicar), entonces podremos afirmar que la fiabilidad de las puntuaciones del test no es generalizable a las diferentes poblaciones y contextos representados en el meta-análisis. Este análisis se lleva a cabo aplicando el estadístico Q de heterogeneidad, que se obtiene mediante (Hedges y Olkin, 1985):

$$Q = \sum_i w_i \left(r_i - r_+\right)^2 \quad , (5)$$

donde r_i puede ser sustituido por cualquiera de las transformaciones antes mencionadas (T_i, M_i, Z_i). Bajo la hipótesis de homogeneidad de los coeficientes de fiabilidad, el estadístico Q se distribuye según chi-cuadrado de Pearson con $k - 1$ grados de libertad. Un resultado estadísticamente significativo para Q indica que los coeficientes de fiabilidad no son generalizables a lo largo de las aplicaciones del test.

(3) *Búsqueda de variables moderadoras.*

Una de las principales potencialidades de los estudios GF es la búsqueda de características de los estudios que puedan explicar la heterogeneidad manifestada por los coeficientes de fiabilidad obtenidos en las aplicaciones del test. Con este propósito se aplican técnicas de análisis de varianza o de regresión, tomando los coeficientes de fiabilidad (o sus transformaciones) como la variable dependiente y las características de los estudios como variables independientes o predictoras. Los análisis estadísticos pueden abordarse sin ningún tipo de ponderación (serían los análisis estadísticos convencionales que se aplican en los estudios empíricos), o bien ponderando por la inversa de la varianza de cada coeficiente de fiabilidad. En este último caso, no existe consenso actualmente sobre si debe asumirse un modelo de efectos fijos o de efectos mixtos (Beretvas y Pastor, 2003; Hedges y Vevea, 1998; Sánchez-Meca, Marín-Martínez y Huedo, 2006). En el modelo de efectos fijos se asume que, para cada nivel de la variable moderadora en cuestión, los coeficientes de fiabilidad obtenidos en las diferentes aplicaciones del test estiman un coeficiente de fiabilidad paramétrico común a todos ellos. En este caso el factor de ponderación queda definido por la inversa de la varianza intra-estudio de cada coeficiente de fiabilidad. En el modelo de efectos mixtos, por el contrario, se asume que para cada nivel de la variable moderadora los coeficientes de fiabilidad proceden de una distribución de coeficientes de fiabilidad paramétricos que difieren entre si debido a otros factores y características de los estudios. En este caso, el factor de ponderación está en función de la varianza intra-estudio de cada coeficiente de fiabilidad y de la varianza inter-estudios estimada a través del conjunto de todos los coeficientes de fiabilidad.

UNA REVISIÓN DE LOS ESTUDIOS DE GENERALIZACIÓN DE LA FIABILIDAD

La falta de consenso sobre el modo de analizar estadísticamente los coeficientes de fiabilidad para comprobar si la fiabilidad de las puntuaciones de un test es generalizable a través de sus múltiples aplicaciones, puede atentar contra la comparabilidad de los resultados

TABLA 1

CARACTERÍSTICAS DE LOS 49 ESTUDIOS GF REVISADOS

CARACTERÍSTICA DEL ESTUDIO	Frecuencia	Porcentaje
Fecha de publicación		
1998-1999	2	4.1
2000-2001	9	18.4
2002-2003	17	34.6
2004-2005	9	18.4
2006-2007	10	20.4
2008	2	4.1
Tipo de coeficiente de fiabilidad calculado		
Consistencia interna	46	93.9
Test-retest	20	40.8
Interjueces (intercodificadores)	6	12.2
Formas paralelas	1	2.0
Otros	4	8.2
¿Se mezclaron distintos coeficientes en un mismo análisis?		
No	44	89.8
Sí	5	10.2
¿Se incluyó algún factor de ponderación?		
No	38	77.6
Sí, ponderando por N	10	20.4
Sí, ponderando por otro factor	1	2.0
Índice empleado en los análisis		
El propio coeficiente de fiabilidad, r	39	79.6
La raíz cuadrada de 1 – r	1	2.0
El índice de fiabilidad, $\sqrt{r}$	1	2.0
La transformación Z de Fisher sobre r	12	24.5
La transformación Z de Fisher sobre $\sqrt{r}$	3	6.1
Otro	2	4.1
¿Cuántos instrumentos se analizaron?		
Sólo uno	41	83.7
Varios	8	16.3
Fuente de publicación		
EPM[a]	34	69.4
Otra	15	30.6

[a] EPM: *Educational and Psychological Measurement*.

obtenidos en diferentes estudios GF. Con objeto de examinar en qué grado los estudios GF difieren en la metodología aplicada y, en particular, en los métodos estadísticos utilizados para la integración cuantitativa de los coeficientes de fiabilidad, realizamos una revisión sistemática de los estudios GF que hasta la fecha se han publicado.

Para localizar los estudios GF realizamos una búsqueda en las bases electrónicas PsycInfo, Eric y Medline, de forma que apareciera en el título o en el abstract del artículo el término "reliability generalization". Completamos la búsqueda mediante la consulta directa de los números de la revista *Educational and Psychological Measurement*, ya que es ésta la revista que más ha promocionado la realización de este tipo de estudios. La búsqueda abarcó los años 1998 (fecha en que se acuñó el término "generalización de la fiabilidad" para referirse a este tipo de estudios) y 2008, y tuvo lugar en junio de 2008.[1] Como resultado de nuestra búsqueda localizamos 49 estudios GF, todos ellos publicados en inglés.

Las características de los estudios RG que registramos para su análisis fueron: (a) el año de publicación, (b) la revista en la que se publicó el estudio, (c) si el estudio RG se centró en un único test o en varios, (d) el tipo de coeficiente de fiabilidad integrado (consistencia interna, test-retest, formas paralelas, dos mitades, intercodificadores, etc.), (e) si en los análisis estadísticos se mezclaron coeficientes de fiabilidad diferentes (por ejemplo, coeficientes alfa de Cronbach con coeficientes de fiabilidad test-retest), (f) el número de artículos y de coeficientes de fiabilidad acumulados en el estudio RG, (g) si los análisis estadísticos se efectuaron con los propios coeficientes de fiabilidad o si se transformaron éstos por alguno de los métodos anteriormente descritos (índice de fiabilidad, Z de Fisher, transformación de la raíz cúbica, T_i, transformación de la raíz cuadrada, M_i, u otros) y (h) si en los análisis estadísticos se ponderó o no por algún factor, tal como el tamaño de la muestra.

Los resultados de nuestra revisión de los estudios RG se presentan en la Tabla 1. Un primer dato digno de mención es que casi el 70% de los estudios GF (34 estudios) se han publicado en la revista *Educational and Psychological Measurement*. Los restantes 15 estudios RG se han publicado en 15 revistas diferentes dentro del ámbito de la psicología y la educación, lo que refleja la amplia difusión que ha alcanzado este tipo de estudios.

[1] En noviembre de 2008 repetimos la búsqueda para localizar posibles estudios de generalización de la fiabilidad que se hubieran publicado en el año 2008 posteriormente a junio y no encontramos ningún estudio adicional.

Los años 2002 y 2003 han sido los más prolíficos en la publicación de estudios RG, con 17 estudios (34.6%), si bien desde el año 2000 en adelante se vienen publicando estudios RG de forma ininterrumpida, a razón de unos 10 estudios cada dos años. La mayoría de los estudios RG se centraron en el análisis de la fiabilidad de un solo test (41 estudios, 83.7%), si bien unos pocos analizaron la generalización de la fiabilidad de varios tests que medían un mismo constructo. Tal es el caso, por ejemplo, del estudio de Dunn, Smith y Montoya (2006), que analizaron la fiabilidad de varios tests que miden la competencia multicultural. Un aspecto en el que se observa una gran variabilidad entre los estudios GF es el número de artículos y de coeficientes de fiabilidad integrados. En concreto, el número de artículos osciló entre los 6 artículos del estudio de Mji y Alkhateeb (2005) sobre el Cuestionario de Concepciones de las Matemáticas, hasta los 138 artículos integrados en el estudio de Bachner y O'Rourke (2007) sobre la entrevista para evaluar la sobrecarga de trabajo de Zarit aplicada a los cuidadores de enfermos. Pero el estudio RG que integró más coeficientes de fiabilidad fue el de Viswesvaran y Ones (2003), con 1359 coeficientes de fiabilidad de los tests de personalidad que evalúan los cinco grandes rasgos de la personalidad. Tanto el número de artículos como el de coeficientes de fiabilidad acumulados en los estudios GF exhiben una distribución asimétrica positiva, con un valor mediano de 38 artículos y 67 coeficientes de fiabilidad, respectivamente.

Pasando a la descripción de cómo varían entre si los estudios GF según la metodología aplicada, en la Tabla 1 se muestra cómo los coeficientes de fiabilidad que con más frecuencia se utilizan en la literatura son los basados en la consistencia interna del test y, en particular, el coeficiente alfa de Cronbach, ya que 46 de los 49 estudios GF analizados (el 93.9%) integraron este tipo de coeficientes. A gran distancia les siguen los coeficientes de fiabilidad por estabilidad temporal (test-retest), con 20 estudios (40.8%), coeficientes de fiabilidad interjueces, con sólo 6 estudios RG (12.2%), un único estudio RG incluyó coeficientes de fiabilidad por formas paralelas (2%) y 4 estudios RG incluyeron otros coeficientes de fiabilidad (8.2%).[2]

Dentro de la variabilidad en los modos de analizar estadísticamente los datos de un estudio RG, los autores sí parecen haber alcanzado un consenso respecto de lo inadecuado que resulta mezclar en una misma integración meta-analítica coeficientes de fiabilidad procedentes de concepciones diferentes de la misma (consistencia interna, estabilidad temporal, etc.). En nuestra revisión hemos encontrado 5 estudios RG que mezclaron en sus análisis estadísticos coeficientes alfa de Cronbach con coeficientes test-retest, de los cuales tres fueron realizados por la propia precursora de este tipo de estudios sobre las escalas MMPI (Vacha-Haase, 1998; Vacha-Haase, Kogan, Tani y Woodall, 2001; Vacha-Haase et al., 2001). Los otros dos estudios que cayeron en esta práctica desaconsejable fueron los de Capraro, Capraro y Henson (2001) sobre la escala de ansiedad a las matemáticas y Caruso (2000) sobre las escalas de personalidad NEO.

Los dos aspectos en los que se observa una mayor heterogeneidad en las prácticas de los estudios RG son, en primer lugar, si se utilizaron los propios coeficientes de fiabilidad para los análisis estadísticos o bien se transformaron, y en segundo lugar, si los coeficientes se ponderaron o no. Respecto de la primera cuestión, encontramos que la práctica más frecuente consiste en utilizar los coeficientes de fiabilidad directamente sin transformarlos. Esto ocurrió en 39 de los 49 estudios RG revisados (79.6%). El segundo modo de análisis más frecuente fue el uso de la transformación a Z de Fisher de los coeficientes de fiabilidad, con 12 estudios RG (24.5%). De estos 12 estudios, 11 de ellos aplicaron la transformación Z de Fisher a coeficientes alfa de Cronbach, a pesar de que esta práctica no es apropiada desde un punto de vista teórico, ya que el coeficiente alfa no es un coeficiente de correlación de Pearson. En menor medida, se ha utilizado la transformación Z de Fisher sobre el índice de fiabilidad en lugar de sobre el coeficiente de fiabilidad (3 estudios RG), un estudio RG utilizó el índice de fiabilidad directamente para la integración meta-analítica y otro estudio utilizó la transformación basada en la raíz cuadrada de $1 - r$. Por tanto, se hace patente que, dentro de la variabilidad analítica, destacan el uso del propio coeficiente de fiabilidad sin transformar y de la transformación Z de Fisher. Resulta paradójico que ningún estudio RG haya utilizado la transformación basada en la raíz cúbica, propuesta por Hakstian y Whalen (1976), a pesar de que ésta es la propuesta más aconsejada por los teóricos del enfoque GF cuando los coeficientes de fiabilidad son alfas de Cronbach y siendo este tipo de coeficientes el que mayoritariamente se integra en los estudios GF.

Respecto de la cuestión de si se ponderaron los análisis estadísticos o no, la práctica más frecuente en los estudios GF es no ponderar los coeficientes de fiabilidad, con 38 de los 49 estudios RG (77.6%). En este caso, los estudios RG aplican técnicas estadísticas con-

[2] Téngase en cuenta que la suma de todos los porcentajes supera el 100% porque algunos estudios GF registraron coeficientes de fiabilidad de varios tipos.

vencionales, tales como el cálculo de la media aritmética para obtener una estimación global de la fiabilidad de las puntuaciones del test, y pruebas *t* de diferencias entre medias, análisis de varianza y análisis de regresión para explorar el posible influjo de características de los estudios sobre las estimaciones de la fiabilidad. Sólo 11 estudios RG analizaron estadísticamente los coeficientes de fiabilidad aplicando algún factor de ponderación. De los 11 estudios 10 ponderaron por el tamaño de la muestra y sólo uno de ellos ponderó por la inversa de la varianza de cada coeficiente de fiabilidad (Beretvas, Suizzo, Durham y Yarnell, 2008). Por último, el uso de los modelos de efectos aleatorios y mixtos cuando se aplica algún método de ponderación está infrautilizado, ya que tan sólo los dos estudios GF realizados por Beretvas y sus colegas (Beretvas, Meyers y Leite, 2002; Beretvas et al., 2008) han aplicado este tipo de modelos en los que la ponderación por la inversa de la varianza de cada coeficiente de fiabilidad implica sumar la varianza intra-estudio y una estimación de la varianza inter-estudios. En su defecto, son los modelos de efectos fijos los que dominan el panorama de los estudios GF cuando se ponderan los coeficientes de fiabilidad.

CONCLUSIÓN

El propósito de este artículo fue ofrecer una visión panorámica de un método meta-analítico relativamente reciente denominado "generalización de la fiabilidad", que posibilita comprobar empíricamente si las estimaciones de la fiabilidad obtenidas con las múltiples aplicaciones de un test a lo largo de diferentes muestras y contextos es generalizable o si, por el contrario, la fiabilidad de las puntuaciones de un test es específica de las características de la muestra en cuestión y de las condiciones de aplicación del mismo. Los estudios GF permiten comprobar qué características de los estudios empíricos, de las muestras empleadas y de las condiciones de aplicación del test afectan a las estimaciones de la fiabilidad. Además, hemos presentado las fases en las que se lleva a cabo un estudio GF y cuáles son los aspectos estadísticos y psicométricos de este enfoque que actualmente son objeto de estudio y discusión.

En la base del enfoque de la generalización de la fiabilidad se encuentra la crítica, planteada desde hace tiempo por los psicómetras y agudizada en la última década, contra la creencia errónea y muy difundida entre los investigadores y los profesionales de la psicología y de todas aquellas disciplinas en las que se aplican tests psicológicos, educativos o de índole similar, 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 0.80", son incorrectas. Lo correcto es decir "la fiabilidad de las puntuaciones del test en esta muestra es 0.80". Esa práctica, bautizada recientemente como 'inducción de la fiabilidad', de asumir como propia la fiabilidad obtenida en una aplicación previa del test es, en el mejor de los casos, arriesgada, ya que sólo es válida si las características de la muestra de sujetos en cuestión y las condiciones de aplicación del test son similares a las de la aplicación previa.

Los estudios GF están demostrando empíricamente que la fiabilidad de las puntuaciones de un test varía sensiblemente de una aplicación a otra del mismo y, en consecuencia, los investigadores en psicología en particular y en ciencias sociales y de la salud en general, 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 mismo.

No obstante, la revisión de los 49 estudios GF que hasta la fecha se han publicado pone en evidencia la variabilidad metodológica que existe a la hora de analizar estadísticamente los datos de este tipo de estudios, e incluso la existencia de errores serios en la aplicación de dicha metodología. La causa de esta heterogeneidad en las prácticas de los estudios GF hay que buscarla, por una parte, en la recomendación de los propios precursores de los estudios GF de no considerar monolíticamente este tipo de estudios (Vacha-Haase, 1998; Henson y Thompson, 2002) y, por otra, en la propia juventud de este método de integración meta-analítico. En un futuro próximo, los investigadores en la metodología meta-analítica deberían indagar en las prácticas más apropiadas para meta-analizar coeficientes de fiabilidad y clarificar: (a) si es preferible utilizar en los análisis estadísticos los propios coeficientes de fiabilidad o si es mejor transformarlos para normalizar la distribución muestral y estabilizar las varianzas, (b) si es más apropiado ponderar los estudios por el tamaño muestral o por algún otro factor, o si es preferible no ponderarlos, y (c) cuáles son las bondades de los modelos de efectos aleatorios y mixtos frente al modelo de efectos fijos por el momento imperante en este tipo de estudios. En cualquier caso, aunque el enfoque de la generalización de la fiabilidad tiene también sus detractores (Dimitrov, 2002; Sawilowski, 2000a, 2000b) a pesar de esta necesidad de depuración del modo de hacer estudios GF, no cabe duda del 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 esti-

marse con los datos de las propias muestras y evitar inducciones que pueden provocar serios errores en la estimación de la precisión de nuestras medidas.

REFERENCIAS

(Los estudios precedidos por un asterisco fueron incluidos en nuestra revisión de estudios GF.)

*Bachner, Y. G. y O'Rourke, N. (2007). Reliability generalization of responses by care providers to the Zarit Burden Interview. *Aging and Mental Health, 11*, 678-685.

*Barnes, L. L. B., Harp, D. y Jung, W. S. (2002). Reliability generalization of scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement, 62*, 603-618.

*Beretvas, S. N., Meyers, J. L. y Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement, 62*, 570-589.

Beretvas, S. N. y Pastor, D. A. (2003). Using mixed-effects models in reliability generalization studies. *Educational and Psychological Measurement, 63*, 75-95.

*Beretvas, S. N., Suizzo, M.-A., Durham, J. A. y Yarnell, L. M. (2008). A reliability generalization study of scores on Rotter's and Nowicki-Strickland's locus of control scales. *Educational and Psychological Measurement, 68*, 97-119.

*Campbell, J. S., Pulos, S., Hogan, M. y Murry, F. (2005). Reliability generalization of the Psychopathy Checklist applied in youthful samples. *Educational and Psychological Measurement, 65*, 639-656.

*Capraro, R. M. y Capraro, M. M. (2002). Myers-Briggs type indicator score reliability across studies: A meta-analytic reliability generalization study. *Educational and Psychological Measurement, 62*, 590-602.

*Capraro, M. M., Capraro, R. M. y Henson, R. K. (2001). Measurement error of scores on the Mathematics Anxiety Rating Scale across studies. *Educational and Psychological Measurement, 61*, 373-386.

*Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement, 60*, 236-254.

*Caruso, J. C. y Edwards, S. (2001). Reliability generalization of the Junior Eysenck Personality Questionnaire. *Personality and Individual Differences, 31*, 173-184.

*Caruso, J. C., Witkiewitz, K., Belcourt-Dittloff, A. y Gottlieb, J. D. (2001). Reliability scores from the Eysenck Personality Questionnaire: A reliability generalization study. *Educational and Psychological Measurement, 61*, 675-689.

*Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *Journal of General Psychology, 130*, 290-304.

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

Cooper, H. y Hedges, L. V. (Eds.) (1994). *The handbook of research synthesis*. Nueva York: Russell Sage Foundation.

*De Ayala, R. J., Vonderharr-Carlson, D. J. y Kim, D. (2005). Assessing the reliability of the Beck Anxiety Inventory scores. *Educational and Psychological Measurement, 65*, 742-756.

*Deditius-Island, H. K. y Caruso, J. C. (2002). An examination of the reliability of scores from Zuckerman's Sensation Seeking Scales, Form V. *Educational and Psychological Measurement, 62*, 728-734.

*Dierdoff, E. C. y Wilson, M. A. (2003). A meta-analysis of job analysis reliability. *Journal of Applied Psychology, 88*, 635-646.

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement, 62*, 783-801.

*Dunn, T. W., Smith, T. B. y Montoya, J. A. (2006). Multicultural competency instrumentation: A review and analysis of reliability generalization. *Journal of Counseling and Development, 84*, 471-482.

Feldt, L. S. y Brennan, R. L. (1989). Reliability. En R. L. Linn (Ed.), *Educational measurement* (3ª ed., pp. 105-146). Nueva York: American Council on Education and Macmillan.

Feldt, L. S. y Charter, R. A. (2006). Averaging internal consistency reliability coefficients. *Educational and Psychological Measurement, 66*, 215-227.

*Graham, J. M., Liu, Y. J. y Jeziorski, J. L. (2006). The Dyadic Adjustment Scale: A reliability generalization meta-analysis. *Journal of Marriage and Family, 68*, 701-717.

Gronlund, N. E. y Linn, R. L. (1990). *Measurement and evaluation in teaching* (6ª ed.). Nueva York: Macmillan.

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

Hall, S. M. y Brannick, M. T. (2002). Comparison of two random-effects methods of meta-analysis. *Journal of Applied Psychology, 87*, 377-389.

*Hanson, W. E., Curry, K. T. y Bandalos, D. L. (2002). Reliability generalization of Working Alliance Inventory Scale scores. *Educational and Psychological Measurement, 62*, 659-673.

Hedges , L. V. y Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V. y Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

Heldref Foundation (1997). Guidelines for contributors. *Journal of Experimental Education, 65*, 95-96.

*Hellman, C. M., Fuqua, D. R. y Worley, J. (2006). A reliability generalization study on the Survey of Perceived Organizational Support: The effects of mean age and number of items on score reliability. *Educational and Psychological Measurement, 66*, 631-642.

*Helms, J. E. (1999). Another meta-analysis of the White Racial Identity Attitude Scale's Cronbach alphas: Implications for validity. *Measurement and Evaluation in Counseling and Development, 32*, 122-137.

*Henson, R. K. y Hwang, D.-Y. (2002). Variability and prediction of measurement error in Kolb's Learning Style Inventory scores: A reliability generalization study. *Educational and Psychological Measurement, 62*, 712-727.

*Henson, R. K., Kogan, L. R. y Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement, 61*, 404-420.

Henson, R. K. y Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "reliability generalization" studies. *Measurement and Evaluation in Counseling and Development, 35*, 113-126.

Hunter, J. E. y Schmidt, F. S. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2ª ed.). Thousand Oaks, CA: Sage.

*Kieffer, K. M., Cronin, C. y Fister, M. C. (2004). Exploring variability and sources of measurement error in Alcohol Expectancy Questionnaire reliability coefficients: A meta-analytic reliability generalization study. *Journal of Studies on Alcohol, 65*, 663-671.

*Kieffer, K. M. y Reese, R. J. (2002). A reliability generalization study of the Geriatric Depression Scale. *Educational and Psychological Measurement, 62*, 969-994.

*Lane, G. G., White, A. E. y Henson, R. K. (2002). Expanding reliability generalization methods with KR-21 estimates: An RG study on the Coopersmith Self-esteem Inventory. *Educational and Psychological Measurement, 62*, 685-711.

*Leach, L. F., Henson, R. K., Odom, L. R. y Cagle, L. S. (2006). A reliability generalization study of the Self-Description Questionnaire. *Educational and Psychological Measurement, 66*, 285-304.

*Li, A. y Bagger, J. (2007). The Balanced Inventory of Desirable Responding (BIDR): A reliability generalization study. *Educational and Psychological Measurement, 67*, 525-544.

*López-Pina, J. A., Sánchez-Meca, J. y Rosa-Alcázar, A. I. (en prensa). The Hamilton Rating Scale for Depression: A reliability generalization study. *International Journal of Clinical and Health Psychology*.

Mason, C., Allam, R. y Brannick, M. T. (2007). How to meta-analyze coefficient-of-stability estimates: Some recommendations based on Monte Carlo studies. *Educational and Psychological Measurement, 67*, 765-783.

*Miller, C. S., Shields, A. L., Campfield, D., Wallace, K. A. y Weiss, R. D. (2007). Substance use scales of the Minnesota Multiphasic Personality Inventory: An exploration of score reliability via meta-analysis. *Educational and Psychological Measurement, 67*, 1052-1065.

*Mji, A. y Alkhateeb, H. M. (2005). Combining reliability coefficients: Toward reliability generalization of the Conceptions of Mathematics Questionnaire. *Psychological Reports, 96*, 627-634.

*Nilsson, J. E., Schmidt, C. K. y Meek, W. D. (2002). Reliability generalization: An examination of the Career Decision-Making Self-efficacy Scale. *Educational and Psychological Measurement, 62*, 647-658.

Onwuegbuzie, A. J. y Daniel, L. G. (2004). Reliability generalization: The importance of considering sample specificity, confidence intervals, and subgroup differences. *Research in the Schools, 11*, 60-71.

*O´Rourke, N. (2004). Reliability generalization of responses by care providers to the Center for Epidemiologic Studies-Depression Scale. *Educational and Psychological Measurement, 64*, 973-990.

*Reese, R. J., Kieffer, K. M. y Briggs, B. K. (2002). A reliability generalization study of select measures of adult attachment style. *Educational and Psychological Measurement, 62*, 619-646.

*Rexrode, K. R., Petersen, S. y O'Toole, S. (2008). The Ways of Coping Style Scale: A reliability generalization study. *Educational and Psychological Measurement, 68*, 262-280.

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

*Ross, M. E., Blackburn, M. y Forbes, S. (2005). Reliability generalization of the patterns of adaptive learning survey goal orientation scales. *Educational and Psychological Measurement, 65*, 451-464.

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

*Ryngala, D. J., Shields, A. L. y Caruso, J. C. (2005). Reliability generalization of the Revised Children's Manifest Anxiety Scale. *Educational and Psychological Measurement, 65*, 259-271.

Sánchez-Meca, J. (2003). La revisión del estado de la cuestión: El meta-análisis. En C. Camisón, M. J. Oltra y M. L. Flor (Eds.), *Enfoques, problemas y métodos de investigación en economía y dirección de empresas* (pp. 101-110). Castellón: ACEDE/Fundació Universitat Jaime I–Empresa.

Sánchez-Meca, J. y Ato, M. (1989). Meta-análisis: Una alternativa metodológica a las revisiones tradicionales de la investigación. En J. Arnau y H. Carpintero (Coords.), *Tratado de psicología general I: Historia, teoría y método* (pp. 617-669). Madrid: Alhambra.

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

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

Sánchez-Meca, J. y Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods, 13*, 31-48.

Sánchez-Meca, J., Marín-Martínez, F. y Huedo, T. (2006). Modelo de efectos fijos versus modelo de efectos aleatorios. En J. L. R. Martín, A. Tobías y T. Seoane (Coords.), *Revisiones Sistemáticas en Ciencias de la Vida* (pp. 189-204). Toledo: FISCAM.

Sawilowsky, S. S. (2000a). Psychometrics versus datametrics: Comment on Vacha-Haase's 'Reliability generalization' method and some *EPM* editorial policies. *Educational and Psychological Measurement, 60*, 157-173.

Sawilowsky, S. S. (2000b). Reliability: Rejoinder to Thompson and Vacha-Haase. *Educational and Psychological Measurement, 60*, 196-200.

*Shields, A. L. y Caruso, J. C. (2003). Reliability generalization of the Alcohol Use Disorders Identification Test. *Educational and Psychological Measurement, 63*, 404-413.

*Shields, A. L. y Caruso, J. C. (2004). A reliability induction and reliability generalization study of the Cage Questionnaire. *Educational and Psychological Measurement, 64*, 254-270.

*Shields, A. L., Howell, R. T., Potter, J. S. y Weiss, R. D. (2007). The Michigan Alcoholism Screening Test and its shortened form: A meta-analytic inquiry into score reliability. *Substance Use and Misuse, 42*, 1-18.

Silver , N. y Dunlap, W. (1987). Averaging coefficients: Should Fisher's z-transformation be used? Journal of Applied Psychology, 72, 3-9.

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

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

*Thompson, B. y Cook, C. (2002). Stability of the reliability of LibQUAL+TM scores: A reliability generalization meta-analysis study. *Educational and Psychological Measurement, 62*, 735-743.

Thompson, B. y Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174-195.

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., 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.

*Vacha-Haase, T., Kogan, L. R., Tani, C. R. y Woodall, R. A. (2001). Reliability generalization: Exploring variation of reliability coefficients of MMPI clinical scales scores. *Educational and Psychological Measurement, 61*, 45-59.

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.

Vacha-Haase, T. y Ness, C. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *Journal of Experimental Education, 67*, 335-342.

*Vacha-Haase, T., Tani, C. R., Kogan, L. R., Woodall, R. A. y Thompson, B. (2001). Reliability generalization: Exploring reliability variations on MMPI/MMPI-2 validity scale scores. *Assessment, 8*, 391-401.

*Visweswaran, C. y Ones, D. S. (2000). Measurement error in 'big five factors' personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement, 60*, 224-235.

*Voskuijl, O. F. y van Sliedregt, T. (2002). Determinants of interrater reliability of job analysis: A meta-analysis. *European Journal of Psychological Assessment, 18*, 52-62.

*Wallace, K. A. y Wheeler, A. J. (2002). Reliability generalization of the Life Satisfaction Index. *Educational and Psychological Measurement, 62*, 674-684.

Wilkinson, L. & APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement, 58*, 21-37.

*Yin, P. y Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60*, 201-223.

*Youngstrom, E. A. y Green, K. W. (2003). Reliability generalization of self-report of emotions when using the Differential Emotions Scale. *Educational and Psychological Measurement, 63*, 279-295.

*Zangaro, G. A. y Soeken, K. L. (2005). Meta-analysis of the reliability and validity of Part B of the Index of Work Satisfaction across studies. *Journal of Nursing Measurement, 13*, 7-22.

NOTAS DE AUTOR

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