EL ENFOQUE META-ANALÍTICO DE GENERALIZACIÓN DE LA FIABILIDAD[1]

THE META-ANALYTIC APPROACH OF RELIABILITY GENERALIZATION

SÁNCHEZ-MECA, J. Y LÓPEZ-PINA, J. A.
Universidad de Murcia

Resumen

Frases del tipo «la fiabilidad del test es 0.80» son incorrectas. Es más apropiado decir «la fiabilidad de las puntuaciones del test en una determinada aplicación del mismo es 0.80». El enfoque meta-analítico de generalización de la fiabilidad 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 de generalización de la fiabilidad: (a) formulación del problema, (b) búsqueda de los estudios, (c) codificación de los estudios, (d) análisis estadístico e interpretación y (e) publicación. Se presenta una visión actualizada de los problemas estadísticos de este enfoque: (a) transformar versus no transformar los coeficientes, (b) ponderar versus no ponderar los coeficientes, (c) cómo tratar la dependencia estadística entre los coeficientes y (d) cuál es el modelo estadístico más apropiado (efectos fijos, efectos aleatorios, efectos mixtos).

Abstract

Sentences such as «the test reliability is 0.80» are wrong. It is more appropriate to say «the test scores reliability in a given application of it is 0.80». The meta-analytic approach of reliability generalization 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 a reliability generalization study are presented: (a) formulating the problem, (b) searching for the studies, (c) coding studies, (d) statistical analysis and interpretation, and (e) publication. An updated overview of the statistical problems of this approach: (a) to transform versus not to transform the reliability coefficients, (b) to weight versus not to weight the coefficients, (c) how to manage statistical dependency among the coefficients, and (d) which statistical model is the most appropriate (fixed-, random-, and mixed-effects).

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

Palabras Clave

Generalización de la fiabilidad; meta-análisis; coeficiente de fiabilidad

Key Words

Reliability generalization; meta-analysis; reliability coefficient

Introducción

Tanto la investigación psicológica como el desempeño profesional en cualquier ámbito de la psicología requieren del uso de tests psicológicos y de otros instrumentos de evaluación y medición. Conocer la calidad métrica de estos instrumentos de medida es fundamental para su correcto uso e interpretación. De entre las distintas propiedades psicométricas de los instrumentos de medida, la fiabilidad es una de las más relevantes. Sin embargo, es poco frecuente encontrar en las investigaciones psicológicas que utilizan instrumentos de medida alguna alusión a la fiabilidad de sus datos. Aunque sí es frecuente encontrar en las investigaciones alusiones a la fiabilidad de un instrumento obtenida en una aplicación previa del mismo, ya sea la aportada en el manual del test o la estimada en investigaciones previas por los propios investigadores o por otros.

Esa práctica de aludir a la fiabilidad del test obtenida en aplicaciones previas del mismo se basa en una interpretación errónea del concepto de fiabilidad que está muy arraigada entre los investigadores y los psicólogos aplicados. Así, afirmaciones del tipo «el test tiene una fiabilidad de 0,80» es incorrecta. Frases de este tipo se basan en la idea errónea de que la fiabilidad es una propiedad inherente al test y, en consecuencia, inmutable una vez obtenida en la muestra normativa que sirvió para baremarlo. En realidad, la fiabilidad es una propiedad de las puntuaciones obtenidas al aplicar un test a una muestra concreta de sujetos (cf. por ejemplo, Crocker y Algina, 1986; Dawis, 1987; Gronlund y Linn, 1990; Pedhazur y Schmelkin, 19991; Traub, 1994). En palabras de Gronlund y Linn (1990): «Reliability refers to the results obtained with an evaluation instrument and not to the instrument itself. Thus, it is more appropriate to speak of the reliability of 'test scores' or the 'measurement' than of the 'test' or the 'instrument'» (p. 78).

En consecuencia, la fiabilidad de las puntuaciones de un test puede variar sensiblemente en sucesivas aplicaciones del mismo. Como afirma Rowley (1976): «An instrument itself is neither reliable nor unreliable ... A single instrument can produce scores which are reliable, and other scores which are unreliable» (p. 53). Por una parte, las estimaciones de la fiabilidad obtenidas al aplicar un test a diferentes muestras de sujetos, aunque hayan sido seleccionadas al azar de una misma población, variarán entre sí por mero error de muestreo aleatorio. Por otra parte, en la medida en que las muestras pertenezcan a poblaciones de referencia diferentes, ello hará que difieran en cuanto a su composición y a la variabilidad de las puntuaciones obtenidas y, en consecuencia, las estimaciones de la fiabilidad variarán aún más entre sí que por mero error de muestreo aleatorio (Henson y Thompson, 2002; Vacha-Haase, Kogan y Thompson, 2000). Así pues, aludir a la fiabilidad obtenida con las puntuaciones de otra muestra de sujetos sólo estaría justificado si la muestra actual es igual en composición y variabilidad a la anterior. En palabras de Crocker y Algina (1986): «Potential test users need to determine whether reliability estimates reported in test manuals are based on samples similar in composition and variability to the group for whom the test will be used» (p. 144).

Esta práctica, bastante generalizada entre los investigadores, de asumir para su propia muestra la fiabilidad obtenida en alguna aplicación previa del test, ha sido denominada por Vacha-Haase et al. (2000) como *inducción de la fiabilidad* (reliability induction). Aquí, el término 'inducción' hace referencia al hecho de que el investigador atribuye a las puntuaciones de su muestra la fiabilidad obtenida en un caso particular anterior, como si la fiabilidad obtenida en dicho caso fuera generalizable a otros casos particulares futuros. En palabras de Henson y Thompson (2002), la inducción de la fiabilidad «reflects a generalization of a specific instance

(e.g., test manual coefficient) to a more general state of affaire (e.g., future scores from the test)» (p. 114).

Diversos estudios de revisión demuestran que esta práctica de inducir la fiabilidad a partir de los coeficientes de fiabilidad obtenidos en aplicaciones previas del test es muy frecuente. Así, Vacha-Haase y Ness (1999) encontraron que el 23% de los artículos empíricos publicados en tres revistas psicológicas (*Journal of Counseling Psychology*, *Psychology and Aging* y *Professional Psychology: Research and Practice*) inducían la fiabilidad a partir de aplicaciones previas de los instrumentos de medida y sólo el 35.6% de los artículos aportaron alguna estimación de la fiabilidad a partir de los propios datos analizados. En esta misma línea, Whittington (1998) 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 tets. Y Vacha-Haase, Henson y Caruso (2002), en su revisión de 25 estudios de generalización de la fiabilidad encontraron que, en promedio, el 75.6% de los estudios empíricos que utilizan instrumentos de medida indujeron la fiabilidad a partir de aplicaciones previas del test, mientras que sólo el 25.2% de los estudios aportan estimaciones propias de la fiabilidad. Para que dicho proceso inductivo de la fiabilidad tuviera cierta validez, los investigadores tendrían que comprobar que su muestra de sujetos es similar en composición y variabilidad y, desafortunadamente, ésta es una comprobación que raras veces se hace (Vacha-Haase et al. 2002).

Una de las consecuencias más importantes de desatender la fiabilidad de las puntuaciones de un test en una aplicación concreta del mismo, o de inducirla a partir de aplicaciones previas, tiene que ver con la estimación del tamaño del efecto y con la potencia estadística de los contrastes de hipótesis. Como no puede ser de otra forma, una baja fiabilidad en las puntuaciones de la muestra atenúa la estimación del tamaño del efecto y, al mismo tiempo, disminuye la potencia estadística de las pruebas de significación, como así queda reconocido en el informe de la APA Task Force on Statistical Inference (Wilkinson & APA Task Force on Statistical Inference, 1999). Es por ello que el reporte de la fiabilidad obtenida con las puntuaciones de la muestra permite interpretar adecuadamente las estimaciones del tamaño del efecto y los resultados de las pruebas de significación estadística.

En respuesta a las malas prácticas de inducir la fiabilidad de estudios previos o incluso de no informar de ella en absoluto, la comunidad científica está intentando modificar desde finales del siglo pasado este hábito, concienciando a los investigadores de que estimen la fiabilidad de las puntuaciones de los tests que apliquen en sus investigaciones con los propios datos de su muestra. Prueba de ello es la recomendación hecha por la *APA Task Force on Statistical Inference* al respecto: «It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees ... Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric» (Wilkinson & the APA Task Force on Statistical Inference, 1999, p. 596). 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*, 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).

El enfoque de generalización de la fiabilidad

Dado que la fiabilidad es una propiedad inherente a las puntuaciones obtenidas por un test en una aplicación concreta del mismo sobre una determinada muestra de sujetos, los coeficientes de fiabilidad que se pueden obtener en diferentes aplicaciones de un mismo test variarán en función de diversos factores, tales como el error de muestreo, el modo y condiciones de aplicación del test y la composición y variabilidad de la muestra. En consecuencia, estudiar cómo varían los coeficientes de fiabilidad obtenidos en diferentes aplicaciones de un test constituye una tarea científica muy importante, tanto desde un punto de vista teórico como aplicado.

Para abordar esta tarea el meta-análisis es la metodología idónea, 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 (Botella y Gambara, 2002; Cooper, 1998; Cooper y Hedges, 1994; Glass, McGaw y Smith, 1981; Hedges y Olkin, 1985; Hunter y Schmidt, 2004; Martín, Tobías y Seoane, 2006; Petticrew y Roberts, 2006; Sánchez-Meca y Ato, 1989; Schulze, 2004). Aplicado al estudio de la fiabilidad de las puntuaciones de los tests, el meta-análisis permite integrar mediante el uso de técnicas de análisis estadístico, un conjunto de coeficientes de fiabilidad obtenidos al aplicar un test a diferentes muestras de sujetos, para obtener una estimación media de la fiabilidad de las puntuaciones del test, estudiar la variabilidad de los coeficientes de fiabilidad obtenidos en las muestras 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 provocando tal variabilidad.

Aunque tenemos constancia de que este tipo de estudios meta-analíticos ya se estaban haciendo mucho antes,[1] no ha sido hasta la década de 1990 cuando el meta-análisis de coeficientes de fiabilidad ha iniciado su período de máxima popularidad, principalmente desde que Vacha-Haase (1998) acuñara el término *generalización de la fiabilidad* (reliability generalization) para referirse a este tipo de meta-análisis. En un estudio de generalización de la fiabilidad los coeficientes de fiabilidad resultantes de aplicar un test en diferentes muestras de sujetos se toman como la variable dependiente objeto de análisis y diversas características de los estudios actúan como variables predictoras para determinar cuáles de ellas dan mejor cuenta de la variabilidad de las estimaciones de la fiabilidad. Para alcanzar ese propósito se ponen en juego las técnicas de análisis estadístico más apropiadas a tal fin, tales como contrastes de hipótesis, intervalos de confianza, análisis de varianza, análisis de regresión y, en definitiva, el modelo lineal general (Henson y Thompson, 2002; Thompson, 2003; Vacha-Haase, 1998; Vacha-Haase et al., 2002).

Los estudios de generalización de la fiabilidad tienen su origen directo en el enfoque de *generalización de la validez* (validity generalization) desarrollado por J. E. Hunter y F. L. Schmidt en el ámbito de la psicología de las organizaciones, con el objeto de determinar si los coeficientes de validez de un determinado instrumento de medida obtenidos al aplicar dicho instrumento en diferentes contextos y poblaciones de referencia son homogéneos y, por tanto, la validez de dicho instrumento es generalizable o si, por el contrario, los coeficientes de validez difieren tanto entre sí que no es posible generalizar su validez (Hunter y Schmidt, 1990, 2004; Hunter, Schmidt y Jackson, 1982; Schmidt y Hunter, 1977; puede consultarse Sánchez-Meca, 1999, para una presentación en castellano de este enfoque). La única diferencia entre un tipo y otro de estudios está en el tipo de índices estadísticos que se integran cuantitativamente: coeficientes de validez en los estudios de generalización de la validez y coeficientes de fiabilidad en el de los de generalización de la fiabilidad, si bien en este último caso también se ha propuesto integrar los errores típicos de medida.

Los estudios de generalización de la fiabilidad pueden ofrecer una importante contribución al conocimiento teórico de los psicómetras

[1] Por ejemplo, Salgado y Moscoso (1996) hicieron un meta-análisis de coeficientes de fiabilidad inter-codificadores de las pruebas de rendimiento laboral en los estudios de validez sobre selección de personal. Conway, Jako y Goodman (1995) hicieron un meta-análisis en el que integraron coeficientes de fiabilidad inter-codificadores y de consistencia interna de las entrevistas de selección. Yarnold y Mueser (1989) publicaron un meta-análisis sobre la fiabilidad de las medidas del patrón de conducta Tipo A. Parker, Hanson y Hunsley (1988) hicieron una comparación meta-analítica de la fiabilidad (y la validez) de las aplicaciones de los tests MMPI, Rorschach y WAIS. Y Parker (1983) realizó un meta-análisis de la fiabilidad (y la validez) de las puntuaciones del test de Rorschach. Aunque en estos meta-análisis todavía no se hizo mención del término 'generalización de la fiabilidad', pueden considerarse ejemplos de tal tipo de estudios.

hacia una mejor comprensión de la fiabilidad, los factores que influyen sobre ella y su papel en la interpretación de los resultados de las investigaciones que utilizan tests e instrumentos de medida. Por otra parte, este tipo de estudios también puede ser de gran utilidad a los investigadores aplicados y a los administradores de tests en sus investigaciones aplicadas y en la toma de decisiones sobre grupos de personas y sobre individuos concretos (Vacha-Haase et al. 2002). Por último, este tipo de estudios está contribuyendo a concienciar a los investigadores de la importancia de aportar estimaciones de la fiabilidad con los propios datos en lugar de inducir la fiabilidad o de no hacer mención alguna a la misma.

Aunque el enfoque de la generalización de la fiabilidad también tiene sus detractores (Dimitrov, 2002; Sawilowski, 2000a, 2000b), desde su inicio en 1998 hasta la fecha, ya se han publicado más de 40 estudios de generalización de la fiabilidad sobre muy diversos tests psicológicos e instrumentos de medida. Como ejemplos, caben mencionar los estudios de generalización de la fiabilidad 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).

En lo que sigue presentamos una revisión de los procedimientos más usuales para estimar la fiabilidad y una panorámica de cómo se lleva a cabo un estudio de generalización de la fiabilidad, para lo cual desarrollamos las fases que deben seguirse en su realización, nos detenemos en los aspectos analíticos y estadísticos de esta metodología y nos servimos de un ejemplo para ilustrar qué puede ser capaz de ofrecernos un estudio de generalización de la fiabilidad.

Procedimientos para estimar la fiabilidad

Desde la teoría clásica de tests, la estimación de la fiabilidad de las puntuaciones en un grupo se puede realizar a través de tres procedimientos: test-retest, formas paralelas y dos mitades (Crocker y Algina, 1986; Gulliksen, 1987). El procedimiento test-retest requiere la aplicación del mismo test en dos ocasiones diferentes, generalmente, en un intervalo temporal corto. El procedimiento de las formas paralelas requiere la construcción de dos formas del mismo test estrictamente equivalentes (medias y varianzas iguales). Para evaluar la fiabilidad de las puntuaciones en ambos procedimientos se emplea el coeficiente de correlación producto-momento de Pearson. Sin embargo, el procedimiento de las dos mitades (también conocido como consistencia interna) permite obtener un coeficiente de fiabilidad a partir de una única aplicación del test. Este procedimiento consiste en dividir el test en dos partes equivalentes (e.g., pares vs. impares, 1ª mitad vs. 2ª mitad) y calcular el coeficiente de correlación de Pearson entre las dos mitades formadas. El coeficiente de fiabilidad obtenido corresponde, entonces, al test mitad. Para obtener el coeficiente de fiabilidad en el test completo debemos aplicar la ecuación de Spearman Brown para el caso de longitud doble. Si llevamos el procedimiento de las dos mitades hasta el extremo de considerar que cada ítem equivale a un test, podemos emplear un coeficiente de fiabilidad conocido como coeficiente alfa (Cronbach, 1951).

El coeficiente alfa es una expresión del promedio de las covarianzas entre los ítems de un test, cuando las varianzas de error de los ítems son iguales, es decir, cuando los ítems son estrictamente paralelos se basa en un modelo de medida esencialmente tau-equivalente (Raykov, 1997). Si las varianzas de error de los ítems difieren, entonces el coeficiente alfa es más bajo que el promedio de todos los coeficientes de fiabilidad para las dos mitades de un test estimados a través de la ecuación de Spearman-Brown (Cortina, 1993; Miller, 1995). 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, σ_j^2 es la varianza del ítem j, y σ_X^2 es la varianza total.

Fases de un estudio de generalización de la fiabilidad

Dado que un estudio de generalización de la fiabilidad es un tipo de meta-análisis, sus etapas son básicamente las mismas que las que se suelen proponer para los meta-análisis: (1) formulación del problema, (2) búsqueda de los estudios, (3) codificación de los estudios, (4) análisis estadístico e interpretación y (5) publicación del estudio (Botella y Gambara, 2002; Lipsey y Wilson, 2001; Marín-Martínez, Sánchez-Meca, Huedo y Fernández-Guzmán, 2007; Rosenthal, 1991; Sánchez-Meca, 2003).

Formulación del problema

En un estudio de generalización de la fiabilidad el objetivo es evaluar en qué grado se puede generalizar la fiabilidad de las puntuaciones de un determinado test o instrumento de medida o, lo que es lo mismo, examinar la variabilidad de las estimaciones de la fiabilidad obtenidas al aplicar un test en diferentes contextos y a diferentes muestras de sujetos que pueden proceder de diferentes poblaciones de referencia.[2] Identificar las características de los estudios que afectan a los coeficientes de fiabilidad obtenidos en las aplicaciones de un test suele un objetivo prioritario en este tipo de estudios.

El test se puede haber aplicado en distintos contextos, con diferentes fines o propósitos (e.g., diagnóstico de un trastorno, cribado de población general, etc.), puede haber varias versiones diferentes del test (e.g., 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 generalización de la fiabilidad de las puntuaciones obtenidas con el test.

En la decisión sobre si es apropiado realizar un estudio de generalización de la fiabilidad sobre un determinado test deberían tenerse en cuenta, al menos, dos consideraciones (Henson y Thompson, 2002). En primer lugar, debe tratarse de un test de uso suficientemente extendido como para que tenga sentido integrar estimaciones de la fiabilidad de sus aplicaciones. Un test que a penas se ha aplicado en los estudios científicos no tiene interés someterlo a un estudio de generalización de la fiabilidad de sus puntuaciones.

En segundo lugar, debería existir un número razonable de estudios empíricos que han aplicado el test en cuestión y que aportan estimaciones propias de la fiabilidad de las puntuaciones, así como otros datos estadísticos relevantes, de entre los que no debe faltar la variabilidad de las puntuaciones en la muestra. No es posible, sin embargo, indicar un número mínimo de estudios con la información estadística pertinente como criterio para decidir si es apropiado o no realizar un estudio de generalización de la fiabilidad. Los estudios de generalización de la fiabilidad realizados hasta la fecha son muy variables a este respecto. Así, el número de estimaciones de la fiabilidad meta-analizadas puede ser tan bajo como los 18 coeficientes alfa integrados en el estudio de Campbell et al. (2005) sobre el test *Psychopathy Checklist*, y los 813 coeficientes alfa del estudio de Leach, Henson, Odom y Cagle (2006) sobre el test *Self-Description Questionnaire*.

Búsqueda de los estudios

La tarea más ardua de un estudio de generalización de la fiabilidad es la búsqueda y localización de los estudios empíricos que han aplicado el test en cuestión y que pueden haber

[2] Algunos estudios de generalización de la fiabilidad no se han centrado en un único test, sino en un conjunto de instrumentos de medida de uno o de varios constructos, con objeto de evaluar la fiabilidad media alcanzada con las puntuaciones de cada instrumento y compararlas entre ellos. Tal es el caso, por ejemplo, del estudio realizado por Viswesvaran y Ones (2000), en el que meta-analizaron coeficientes de fiabilidad obtenidos con los tests de personalidad que se utilizan habitualmente en selección de personal, agrupándolos en función de los cinco grandes constructos de la personalidad.

aportado alguna estimación de la fiabilidad con los propios datos de la muestra de sujetos. Para abordar correctamente esta etapa el primer paso consiste en definir claramente los criterios de selección de los estudios. Al establecer dichos criterios no podemos olvidar algunos aspectos importantes. En primer lugar, para ser seleccionados los estudios tienen que ser empíricos y grupales, es decir, tienen que haber utilizado una o varias muestras de sujetos y sobre ellas se tiene que haber aplicado el test objeto de estudio. En segundo lugar, si hay varias versiones del test con diferentes longitudes, o bien existen diferentes adaptaciones del mismo a diferentes idiomas, culturas o edades, tenemos que especificar si nuestro estudio de generalización de la fiabilidad 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. En tercer lugar, es preciso también indicar la población o poblaciones de sujetos sobre las que interesa realizar el estudio de generalización de la fiabilidad, ya que las puntuaciones del test no tendrán la misma fiabilidad cuando éste se aplica a muestras clínicas que a muestras procedentes de la población general, o cuando se aplica el test a franjas de edad diferentes. En cuarto lugar, 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. 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 fijados los criterios de selección de los estudios, se tiene que diseñar un plan de búsqueda de los estudios combinando diferentes sistemas de búsqueda. No puede faltar una búsqueda en las *bases de datos electrónicas* al uso (PsycInfo, Medline, ERIC, etc.). Cuando el test en cuestión tiene como objeto evaluar aspectos relacionados con trastornos psicológicos, psicoeducativos o similares, también conviene consultar las bases de datos de la Colaboración Cochrane[3] y de la Colaboración Campbell,[4] que son dos asociaciones internacionales dirigidas a promover la realización de estudios meta-analíticos de alta calidad en el ámbito de la salud, la educación, el trabajo social y la criminología (cf. Petrosino, Boruch, Soydan, Duggan y Sánchez-Meca, 2001; Sánchez-Meca, Boruch, Petrosino y Rosa-Alcázar, 2002; Shadish, Chacón-Moscoso y Sánchez-Meca, 2005). La estrategia de búsqueda más apropiada en estas bases consiste en buscar el nombre del test (o alguna abreviatura del mismo si la tuviera) en el abstract. Es posible que un reducido número de ellos tenga que ser excluido por ser revisiones teóricas o estudios con $N = 1$. Pero esta estrategia nos asegura que la mayoría de los estudios identificados han aplicado el test a una o varias muestras de sujetos. Como complemento a esta estrategia 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 muy probable que otros muchos estudios que han aplicado el test en cuestión no lo mencionen en el abstract. Por tanto, se hace preciso complementar estas estrategias de búsqueda con otras. Una estrategia de búsqueda complementaria a la anterior, y que es más informal, consiste en examinar estudios de revisión en los que sabemos que se habla del

[3] Puede consultarse la página web de esta asociación en: http://www.cochrane.org. En España el Centro Regional Cochrane Iberoamericano tiene su sede en Barcelona, cuya página web puede consultarse en: http://www.cochrane.es.

[4] Puede consultarse la página web de esta asociación en: http://www.campbellcollaboration.org.

test, así como las referencias de los estudios meta-analíticos realizados sobre temas que necesariamente implican el uso de dicho test. Por ejemplo, si el test en cuestión mide el nivel de ansiedad de los sujetos adultos, será conveniente revisar los estudios meta-analíticos que se hayan publicado sobre los trastornos de ansiedad y, más en concreto, sobre la eficacia del tratamiento de los trastornos de ansiedad. Otra estrategia informal consiste en consultar a investigadores expertos en el tema para que nos envíen trabajos en los que han 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.

Los diferentes procedimientos de búsqueda nos permitirán localizar estudios que pueden haber aplicado el test, pero no aseguran que se haya estimado la fiabilidad con los propios datos de la muestra. En consecuencia, el paso siguiente será conseguir todos esos estudios y leerlos para comprobar cuáles cumplen con nuestros criterios de selección. El conjunto final de trabajos incluidos en nuestro estudio de generalización de la fiabilidad 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

Una vez localizados los estudios que cumplen con los criterios de selección, se tiene que elaborar un protocolo de registro de las características de los estudios así como de las estimaciones de la fiabilidad aportadas por dichos estudios. Para una correcta codificación de los estudios se debe elaborar un libro de codificación que recoja todos los aspectos a tener en cuenta en dicho proceso.

La codificación de las características de los estudios permitirá comprobar cuáles de ellas pueden estar afectando a la variabilidad de los coeficientes de fiabilidad obtenidos en las muestras y, en consecuencia, a entender mejor de qué factores depende la fiabilidad de las puntuaciones de ese test.

En primer lugar, *factores metodológicos* en la aplicación del test pueden provocar variabilidad en los coeficientes de fiabilidad obtenidos. Entre esos factores cabe mencionar, por ejemplo, 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.

En segundo lugar, la *procedencia de la muestra de sujetos* también afectará a las estimaciones de la fiabilidad. Dentro de este bloque habría que mencionar todos aquellos aspectos que tienen que ver con la composición de la muestra y la población de referencia a la que pertenece. Por ejemplo, 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. en la muestra.

Un tercer conjunto de características que también pueden provocar variabilidad en los coeficientes de fiabilidad de un mismo test son *de tipo contextual, o circunstancial*, respecto de la aplicación del test. En este caso cabe mencionar el propósito del estudio, distinguiendo entre estudios psicométricos (e.g., estudio de validación de un test, adaptación de un test, etc.) y estudios de naturaleza sustantiva (e.g., estudio predictivo de factores de riesgo de un trastorno, sobre la eficacia de un tratamiento, estudios diagnósticos, etc.). Otros factores contextuales tiene que ver con 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 y, en definitiva, un largo etcétera que dependerá del test en cuestión.

Además de las diferentes características de los estudios, un elemento fundamental en un estudio de generalización de la fiabilidad es la obtención de alguna *estimación de la fiabilidad* con los propios datos de la muestra. A este res-

pecto, hay que tener en cuenta varios aspectos. En primer lugar, cuando el estudio incluya varias muestras con sus correspondientes coeficientes de fiabilidad deberemos recoger todos ellos. Es decir, más que el estudio (o el artículo) la unidad de análisis en un estudio de generalización de la fiabilidad es la muestra. Por tanto, un mismo estudio puede aportar al meta-análisis más de una unidad de análisis.

En segundo lugar, cabe también la posibilidad de que un estudio aporte más de una estimación de la fiabilidad sobre una misma muestra. 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 éstas se meta-analizarán por separado para evitar problemas de dependencia estadística, como abordaremos más adelante. 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 (e.g., consistencia interna, estabilidad temporal, formas paralelas).

Tanto el proceso de codificación de las características de los estudios como el de obtención de los coeficientes de fiabilidad es muy recomendable someterlos a un estudio de fiabilidad con objeto de verificar si ambos procesos se han realizado de forma precisa. Para ello, un procedimiento económico en tiempo y recursos consiste en seleccionar una muestra aleatoria de todos los estudios del meta-análisis y someter la codificación y la obtención de los coeficientes de fiabilidad a un proceso de codificación doble por dos codificadores independientes. Este análisis permite valorar la precisión en la recogida de los datos para el meta-análisis y depurar el libro de codificación.

Análisis estadístico e interpretación

Una vez que disponemos de la base de datos de todos los estudios que aportan estimaciones propias de la fiabilidad, junto con las características de los estudios codificadas, el paso siguiente consiste en analizar estadísticamente los datos. No existe en la actualidad una visión monolítica sobre cómo deben analizarse los datos en un estudio de generalización de la fiabilidad. De hecho, los propios precursores de este enfoque no plantearon pautas concretas (Henson y Thompson, 2002; Thompson, 2003; Vacha-Haase, 1998), y ello ha llevado a que exista una gran diversidad en los análisis estadísticos que se han aplicado en los estudios de generalización de la fiabilidad publicados hasta la fecha.

Las diferentes propuestas difieren en cuanto a (cf. Beretvas y Pastor, 2003; Feldt y Charter, 2006; Henson y Thompson, 2002; Mason, Allam y Brannick, 2007; Onwuegbuzie y Daniel, 2004; Rodríguez y Maeda, 2006): (a) la conveniencia de ponderar o no cada coeficiente de fiabilidad por algún factor, tales 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 (e.g., 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 (e.g., aplicando contrastes de hipótesis convencionales o no convencionales).

No obstante, sí podemos decir que existe un consenso en cuanto al modo de estructurar los análisis 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, si existe heterogeneidad, (4) 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 generalización de la fiabilidad es caracterizar los estudios empíricos que han aplicado el test, es decir, describir las características de las muestras de sujetos 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, o instantánea, de cuál es el estudio

prototípico en el que se ha aplicado el test. Para alcanzar este objetivo, se utilizan las técnicas estadísticas descriptivas y gráficas al uso: distribuciones de frecuencias para las características cualitativas (e.g., la versión del test) y estadísticos descriptivos básicos para las características cuantitativas (e.g., para la variable edad media de la muestra de sujetos se calcula la media, la desviaciones típicas, el mínimo y el máximo). En cuanto a las representaciones gráficas, se pueden utilizar los típicos diagramas de barras o de sectores, histogramas y los elaborados desde el enfoque del análisis exploratorio de datos, tales como el gráfico en tronco y hojas (stem-and-leaf display) o el gráfico de caja (boxplot).

(2) Estimación de la fiabilidad media

A partir de los coeficientes de fiabilidad obtenidos en las muestras de sujetos aportadas por los estudios empíricos, se calcula un coeficiente de fiabilidad medio que reflejará el nivel global medio de fiabilidad obtenido por las aplicaciones del test.

Es importante en este análisis tener en cuenta que no es apropiado mezclar coeficientes de fiabilidad conceptualmente diferentes, como son los coeficientes de consistencia interna (alfa de Cronbach, KR-20, KR-21), los basados en la estabilidad temporal (test-retest) y los centrados en el muestreo del dominio (formas paralelas). Al ser conceptualmente diferentes, estos coeficientes de fiabilidad estiman errores de medida diferentes. Aunque algunos estudios de generalización de la fiabilidad han mezclado coeficientes diferentes en sus análisis (e.g., Capraro, Capraro y Henson, 2001; Caruso, 2000; Vacha-Haase, 1998), es más apropiado obtener estimaciones de la fiabilidad media por separado para cada tipo de coeficiente (Rodríguez y Maeda, 2006).

Por ejemplo, en su estudio de generalización de la fiabilidad del *Psychopathy Checklist*, Campbell et al. (2005) localizaron 21 estudios empíricos que habían aplicado la escala sobre un total de 28 muestras de sujetos independientes y que aportaron 18 coeficientes alfa y 18 coeficientes de correlación intraclase. Al promediar por separado estos coeficientes reportaron una fiabilidad media para los coeficientes alfa de .85 y de .91 para los coeficientes de correlación intraclase.

Otra práctica que debe evitarse es incorporar en los análisis varios coeficientes de fiabilidad obtenidos en una misma muestra, aunque todos ellos sean del mismo tipo (por ejemplo, varios coeficientes alfa calculados sobre una misma muestra en ocasiones diferentes). La inclusión de más de un coeficiente por muestra viola el supuesto de independencia propio de las técnicas meta-analíticas.

Una de las cuestiones que ha generado cierta polémica en el campo de la generalización de la fiabilidad tiene que ver con la conveniencia o no de transformar los coeficientes de fiabilidad, y no parece que existan respuestas definitivas al respecto por el momento. Hay autores que no recomiendan en absoluto transformar los coeficientes de fiabilidad (Hall & Brannick, 2000; Hunter & Schmidt, 2004; Mason et al., 2007), mientras que otros son partidarios de transformarlos (Hedges y Olkin, 1985; Sawilowsky, 2000a; Silver y Dunlap, 1987; Thompson y Vacha-Haase, 2000).

En nuestra opinión, los coeficientes de fiabilidad que se calculan como si fueran coeficientes de correlación de Pearson (e.g., 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. Para ello, aplicaríamos la fórmula de transformación:

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

siendo r_i el coeficiente de fiabilidad estimado en la iésima muestra y Z_i el coeficiente transformado.

Cuando los coeficientes de fiabilidad que se pretende meta-analizar son coeficientes alfa, o similares, no es apropiada la transformación *Z* de Fisher,[5] ya que éstos no se obtienen como si

[5] No obstante, una estrategia que se ha utilizado en algunos estudios de generalización de la fiabilidad con coeficientes alfa consiste en aplicar la transformación *Z* de Fisher sobre el índice de fiabilidad, no sobre el coeficien-

fueran correlaciones de Pearson. En su lugar, se ha propuesto otra transformación consistente en obtener su raíz cúbica mediante la fórmula derivada por Hakstian y Whalen (1976; cf. también Feldt y Charter, 2006; Rodríguez y Maeda, 2006):

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

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 obtiene mediante:

$$r_+ = \frac{\sum_i w_i r_i}{\sum_i w_i}, \quad (3)$$

siendo w_i el factor de ponderación asignado a cada coeficiente de fiabilidad y que suele estar en función del grado de precisión de cada coeficiente. Más adelante abordaremos esta cuestión. Si se ha optado por aplicar alguna transformación de los coeficientes de fiabilidad, Z_i ó T_i, la media se obtiene, respectivamente, mediante:

$$Z_+ = \frac{\sum_i w_i Z_i}{\sum_i w_i}, \quad (4)$$

$$T_+ = \frac{\sum_i w_i T_i}{\sum_i w_i}. \quad (5)$$

Pero Z_+ y T_+ son promedios que no están en la métrica del coeficiente de fiabilidad, por lo que para facilitar su interpretación deben retransformarse mediante las ecuaciones inversas a las ecuaciones (1) y (2):

$$r_+ = \frac{e^{2Z_+} - 1}{e^{2Z_+} + 1} \quad (6)$$

$$r_+ = 1 - T_+^3. \quad (7)$$

Otras opciones, que no detallaremos aquí, consisten en utilizar el índice de fiabilidad en lugar del coeficiente de fiabilidad (es decir, la raíz cuadrada del coeficiente de fiabilidad), ya que conceptualmente éste se define como el cociente entre dos varianzas (la de las puntuaciones verdaderas y la de las puntuaciones empíricas); o también aplicar la transformación Z de Fisher al índice de fiabilidad en lugar de al coeficiente de fiabilidad.

La estimación media de un conjunto de coeficientes de fiabilidad puede hacerse calculando una simple media aritmética de los coeficientes de fiabilidad (e.g., Barnes et al., 2002) o ponderando cada uno de ellos por algún factor de ponderación que refleje su grado de precisión. En el primer caso, las ecuaciones (3), (4) y (5) se simplifican, ya que hacemos $w_i = 1$.

Estudios de simulación Monte Carlo muestran que es más apropiado ponderar los coeficientes de fiabilidad que no ponderarlos, ya que se obtiene una estimación más eficiente de la fiabilidad media (Feldt y Charter, 2006; Mason et al., 2007; Rodríguez y Maeda, 2006). El grado de precisión de un coeficiente de fiabilidad está directamente relacionado con el tamaño muestral, por lo que el tamaño de la muestra es uno de los métodos de ponderación más habitual en los estudios de generalización de la fiabilidad; es decir, en este caso hacemos $w_i = N_i$ en las ecuaciones (3), (4) y (5), siendo N_i el tamaño de la iésima muestra (e.g., Beretvas, Meyers y Leite, 2002). Pero el factor de ponderación que logra la menor varianza de error es el que se obtiene calculando la inversa de la varianza de la distribución muestral del estadístico en cuestión (en nuestro caso, del coeficiente de fiabilidad o de la transformación elegida). Cuando utilizamos la ecuación (3) es porque los coeficientes de fiabilidad que estamos meta-analizando se calculan como correlaciones de Pearson y, en consecuencia, utilizamos como estimador de la varianza muestral de r_i:

---

te de fiabilidad (es decir, sobre la raíz cuadrada del coeficiente de fiabilidad), entendiendo que si el coeficiente de fiabilidad es una proporción de varianzas, al igual que el coeficiente de determinación, su raíz cuadrada permite aproximar el coeficiente alfa a la métrica de un coeficiente de correlación (cf. e.g., Beretvas, Meyers y Leite, 2002).

$$S^2_{r_i} = \frac{\left(1-r_i^2\right)^2}{N_i - 2}. \quad (8)$$

Cuando hemos transformado los coeficientes de fiabilidad [ecuaciones (4) y (5)] las varianzas muestrales de Z_i y de T_i son, respectivamente (Rodríguez y Maeda, 2006):

$$S^2_{Z_i} = \frac{1}{N_i - 3} \quad (9)$$

$$S^2_{T_i} = \frac{18J_i\left(N_i-1\right)\left(1-r_i\right)^{2/3}}{\left(J_i-1\right)\left(9N_i-11\right)^2}, \quad (10)$$

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:

$$w_i = \frac{1}{S^2_{r_i}} = \frac{1}{S^2_{Z_i}} = \frac{1}{S^2_{T_i}}, \quad (11)$$

según que estemos integrando coeficientes de fiabilidad, Z_S de Fisher o transformaciones T, respectivamente. Por tanto, las ecuaciones (3), (4) y (5) para estimar la fiabilidad media de un conjunto de k muestras pueden adoptar distintas formas dependiendo de que no deseemos 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 [ecuación (1)] cuando el coeficiente de fiabilidad en cuestión se calcule como una correlación de Pearson, y utilizar la transformación T [ecuación (2)] 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 (Feldt y Brennan, 1989; Hakstian y Whalen, 1976). 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 (Rodríguez y Maeda, 2006).

Junto con la estimación de la fiabilidad media se suele calcular un *intervalo de confianza* asumiendo una distribución normal. Dependiendo de que utilicemos los valores r_i, Z_i ó T_i, las ecuaciones que nos permiten obtener los intervalos de confianza son, respectivamente:

$$r_+ \pm \left|z_{\alpha/2}\right| S_{r_+} = \begin{cases} r_s = r_+ + \left|z_{\alpha/2}\right| S_{r_+} \\ r_i = r_+ - \left|z_{\alpha/2}\right| S_{r_+} \end{cases} \quad (12)$$

$$Z_+ \pm \left|z_{\alpha/2}\right| S_{Z_+} = \begin{cases} Z_s = Z_+ + \left|z_{\alpha/2}\right| S_{Z_+} \\ Z_i = Z_+ - \left|z_{\alpha/2}\right| S_{Z_+} \end{cases} \quad (13)$$

$$T_+ \pm \left|z_{\alpha/2}\right| S_{T_+} = \begin{cases} T_s = T_+ + \left|z_{\alpha/2}\right| S_{T_+} \\ T_i = T_+ - \left|z_{\alpha/2}\right| S_{T_+} \end{cases}, \quad (14)$$

donde $z_{\alpha/2}$ es la puntuación de la distribución normal estándar para un nivel de significación α; S_{r_+}, S_{Z_+} y S_{T_+}, y son los errores estándar de los índices r_i, Z_i y T_i, respectivamente, que se obtienen por procedimientos diferentes dependiendo del factor de ponderación, w_i, utilizado. En el caso de las ecuaciones (13) y (14), que están en función de los coeficientes transformados a Z de Fisher o a T, será preciso retransformar sus límites confidenciales superior e inferior, con objeto de facilitar su interpretación devolviéndolos a la escala del coeficiente de fiabilidad. Para ello, utilizamos las ecuaciones (6) y (7), respectivamente, sobre los límites confidenciales.

Cuando se calcula la fiabilidad media sin ponderar, es decir, haciendo $w_i = 1$, los errores estándar de los índices r_i, Z_i y T_i se obtienen por el procedimiento convencional de dividir la desviación típica de los coeficientes (S_r, S_Z y S_T) por la raíz cuadrada del número de muestras, k:

$$S_{r_+} = \frac{S_r}{\sqrt{k}} \quad (15)$$

$$S_{Z_+} = \frac{S_Z}{\sqrt{k}} \quad (16)$$

$$S_{T_+} = \frac{S_T}{\sqrt{k}}. \quad (17)$$

Cuando se aplica algún procedimiento de ponderación, ya sea por el tamaño muestral ($w_i = N_i$) o por la inversa de la varianza muestral [ecuación (11)], los errores estándar de los índices r_i, Z_i y T_i, se obtienen mediante la expresión general:

$$S_{r_+} = S_{Z_+} = S_{T_+} = \frac{1}{\sqrt{\sum_i w_i}}. \qquad (18)$$

(3) Evaluación de la heterogeneidad

El elemento clave en el análisis estadístico de un estudio de generalización de la fiabilidad es el que se centra en examinar la variabilidad de los coeficientes de fiabilidad obtenidos con el test en sus diferentes aplicaciones. Si los coeficientes de fiabilidad son homogéneos, entonces podremos afirmar que la fiabilidad de las puntuaciones del test es generalizable a toda la población de estudios empíricos que están representados en el meta-análisis y, en consecuencia, la fiabilidad media estimada mediante cualquiera de los procedimientos presentados en el punto anterior será un buen indicador del nivel general de fiabilidad exhibido por las puntuaciones del test en sus diversas aplicaciones. Si, por el contrario, los coeficientes de fiabilidad son muy heterogéneos entre sí, entonces tendremos que concluir que la fiabilidad de las puntuaciones del test no es generalizable a las diferentes poblaciones y contextos representados en el meta-análisis y, por tanto, la estimación media de la fiabilidad no será de gran utilidad.

Para comprobar el grado de heterogeneidad exhibido por los coeficientes de fiabilidad muestrales se pueden utilizar varias estrategias analíticas. Una primera aproximación consiste en construir algún gráfico que nos dé una idea de la variabilidad exhibida por los coeficientes. Por ejemplo, un histograma de barras, un gráfico en tronco y hojas, un gráfico de caja, etc. Especial interés tiene para este propósito un gráfico denominado 'forest plot', que se ha desarrollado dentro del marco de la metodología meta-analítica, y que consiste en presentar cada coeficiente de fiabilidad muestral con su intervalo de confianza en torno a él mediante un segmento (Henson y Thompson, 2002). Los intervalos de confianza para cada coeficiente de fiabilidad se pueden construir de diferentes formas, según que se trate de correlaciones de Pearson, coeficientes alfa de Cronbach, o de que se aplique alguna transformación (cf. Feldt y Brennan, 1989; Hakstian y Whalen, 1976; Henson, 2001).

Pero el procedimiento más apropiado para determinar si un conjunto de coeficientes de fiabilidad son homogéneos consiste en aplicar el estadístico *Q* de heterogeneidad, que se obtiene mediante:

$$Q = \sum_i w_i \left(r_i - r_+\right)^2, \qquad (19)$$

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 meta-análisis. El factor de ponderación, w_i, tiene que ser necesariamente el que viene definido por la inversa de la varianza del índice en cuestión, y que quedó definido en la ecuación (11). Bajo la hipótesis nula de homogeneidad de los coeficientes de fiabilidad, el estadístico *Q* se distribuye según chi-cuadrado de Pearson con $k - 1$ grados de libertad. Por tanto, asumiendo un nivel de significación (e.g., $\alpha = .05$), podremos rechazar dicha hipótesis de homogeneidad si el estadístico *Q* supera el valor . No obstante, con un número reducido de muestras ($k < 30$) el estadístico *Q* tiene baja potencia estadística para detectar heterogeneidad (Harwell, 1997; Sánchez-Meca y Marín-Martínez, 1997), por lo que es recomendable complementar el resultado del estadístico *Q* con el cálculo del 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 (Higgins y Thompson, 2002; Huedo-Medina, Sánchez-Meca, Marín-Martínez y Botella, 2006). El índice I^2 se obtiene mediante la ecuación:

$$I^2 = \frac{Q - \left(k - 1\right)}{Q} \times 100. \qquad (20)$$

Téngase en cuenta, en primer lugar, que el hecho de multiplicar por 100 tiene como único propósito facilitar la interpretación de este ín-

gundo lugar, que cuando Q es menor que $(k - 1)$, entonces el valor I^2 se iguala a 0 para evitar valores negativos. Higgins y Thompson (2002) propusieron una guía tentativa para ayudar a interpretar este índice, de forma que valores de I^2 en torno a 25%, 50% y 75% pueden interpretarse como reflejando una variabilidad entre los coeficientes de magnitud baja, media y alta. En cualquier caso, un índice $I^2 = 25\%$ ya se puede considerar de magnitud suficiente como para asumir que los coeficientes de fiabilidad muestrales son heterogéneos entre sí, aunque el estadístico Q no haya alcanzado la significación estadística, y que, por tanto, la fiabilidad de las puntuaciones del test no sea generalizable.

Si existe heterogeneidad más allá de la que puede explicar el mero error de muestreo aleatorio, entonces será preciso examinar qué características de los estudios pueden estar afectando a la variabilidad de los coeficientes de fiabilidad. Pero también será apropiado estimar la fiabilidad media exhibida por el conjunto de estudios como un indicador aproximado del nivel medio de fiabilidad obtenido en las aplicaciones del test. El cálculo de la fiabilidad media debería hacerse con alguna de las ecuaciones (3), (4) ó (5) presentadas más arriba, pero el factor de ponderación apropiado no es ninguno de los planteados hasta ahora. Ello se debe a que los promedios arriba propuestos parten del supuesto de que todos los estudios proceden de una misma población general de estudios con un coeficiente de fiabilidad poblacional común a todos ellos. Este modelo estadístico es el que se conoce en el contexto meta-analítico como *modelo de efectos fijos* (Hedges, 1994; Marín-Martínez y Sánchez-Meca, 1998; Sánchez-Meca y Marín-Martínez, 1998). Desde este modelo el factor de ponderación óptimo, w_i, se basa en la inversa de la varianza de la distribución muestral del estadístico definido en la ecuación (11) (ya sea r_i, Z_i ó T_i).

Sin embargo, cuando existe heterogeneidad entre los coeficientes de fiabilidad, el cálculo del efecto medio debe abordarse desde el *modelo de efectos aleatorios*, según el cual, los estudios proceden de una distribución poblacional de coeficientes de fiabilidad paramétricos. Es decir, en lugar de existir un único coeficiente de fiabilidad paramétrico (modelo de efectos fijos), se asume una distribución de coeficientes de fiabilidad paramétricos, a partir de la cual se han ido seleccionando los estudios empíricos del meta-análisis (Hedges y Vevea, 1998; Sánchez-Meca, Marín-Martínez y Huedo, 2006). En consecuencia, el factor de ponderación óptimo implica calcular la inversa de la suma de dos varianzas: la varianza muestral del estadístico en cuestión [según las ecuaciones (8), (9) y (10)] y la varianza inter-estudios, τ^2. Aunque existen diferentes estimadores de la varianza inter-estudios (cf. Sánchez-Meca y Marín-Martínez, 2008; Viechtbauer, 2005, 2007), presentamos aquí uno basado en el método de los momentos y de fácil cálculo propuesto por DerSimonian y Laird (1986):

$$\tau^2 = \frac{Q - \left(k - 1\right)}{c}, \qquad (21)$$

siendo c:

$$c = \sum_i w_i - \frac{\sum_i w_i^2}{\sum_i w_i}. \qquad (22)$$

Cuando Q es menor que $(k - 1)$, τ^2 se trunca al valor 0 para evitar valores negativos. En consecuencia, el factor de ponderación, w_i^{EA}, en un modelo de efectos aleatorios se define como:

$$w_i^{EA} = \frac{1}{S_{r_i}^2 + \tau^2} = \frac{1}{S_{Z_i}^2 + \tau^2} = \frac{1}{S_{T_i}^2 + \tau^2}. \qquad (23)$$

Y la fiabilidad media desde este modelo se obtiene adaptando las ecuaciones (3), (4) y (5) al nuevo factor de ponderación:

$$r_+ = \frac{\sum_i w_i^{EA} r_i}{\sum_i w_i^{EA}} \qquad (24)$$

$$Z_+ = \frac{\sum_i w_i^{EA} Z_i}{\sum_i w_i^{EA}} \qquad (25)$$

$$T_{+} = \frac{\sum_{i} w_i^{EA} T_i}{\sum_{i} w_i^{EA}}. \quad (26)$$

Por último, se puede construir un intervalo de confianza para la fiabilidad media adaptando las ecuaciones (12), (13) y (14) al factor de ponderación de efectos aleatorios.

(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 las variables moderadoras como variables independientes (o predictoras) y los coeficientes de fiabilidad (o su transformación a *Z* ó a *T*) como variable dependiente, se pueden aplicar contrastes de hipótesis, tales como ANOVA cuando la variable independiente es cualitativa (e.g., el idioma en que se aplicó el test) y análisis de regresión cuando es continua (e.g., 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 generalización de la fiabilidad 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*, en cuyo caso el factor de ponderación queda definido por la ecuación (11). 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 inter-estudios, τ^2, actuando la variable moderadora como un factor de efectos fijos. En este caso, la varianza inter-estudios viene estimada por la ecuación:

$$\tau^2 = \frac{Q_E - (k - h)}{tr\mathbf{W} - tr\left[\mathbf{WX}\left(\mathbf{X'WX}\right)^{-1}\mathbf{X'W}\right]} \quad (27)$$

donde *h* es el número de parámetros del modelo; **X** es la matriz de diseño que incluye la(s) variable(s) moderadora(s), **W** es una matriz diagonal que incluye el factor de ponderación para cada coeficiente de fiabilidad; *tr* es la traza de una matriz y Q_E es la suma de cuadrados de error por mínimos cuadrados ponderados, que se obtiene mediante:

$$Q_E = \mathbf{T'WT}, \quad (28)$$

siendo **T** el vector de coeficientes de fiabilidad. Cuando la ecuación (27) da un valor negativo. Éste se trunca al valor 0.

Tanto si asumimos un modelo de efectos fijos con moderadores como uno de efectos mixtos, el ANOVA ponderado que se aplica sobre una variable moderadora cualitativa permite calcular un estadístico, Q_B, que evalúa la existencia de diferencias entre las fiabilidades medias de las diferentes categorías de la variable moderadora. Bajo la hipótesis nula de igualdad de medias, el estadístico Q_B se distribuye según chi-cuadrado con $c - 1$ grados de libertad, siendo *c* el número de categorías de dicha variable. Además, el análisis estadístico se completa con el cálculo del estadístico Q_W, que evalúa si los coeficientes de fiabilidad de cada categoría son homogéneos en torno a su propia fiabilidad media. Bajo la hipótesis nula de homogeneidad de los coeficientes de fiabilidad intra-categoría, el estadístico Q_W se distribuye según chi-cuadrado con $k - c$ grados de libertad. Además, para cada categoría es posible obtener un estadístico, Q_{Wj}, que evalúa la hipótesis de homogeneidad de los coeficientes de fiabilidad para cada categoría por separado. Cada uno de estos estadísticos Q_{Wj} se distribuye según chi-cuadrado de Pearson con $k_j - 1$ grados de libertad, siendo k_j el número de coeficientes de fiabilidad de la categoría *j*.

En el caso de que la variable moderadora sea continua, el modelo de análisis de regresión simple ponderada que se aplicaría, tanto desde el modelo de efectos fijos con moderadores como desde el modelo de efectos mixtos, pasa por calcular el estadístico Q_R, que evalúa si la variable moderadora está estadísticamente asociada a los coeficientes de fiabilidad. Si el modelo es de regresión simple, bajo la hipótesis nula de que la variable moderadora no está asociada a los coeficientes de fiabilidad, el estadístico Q_R se distribuye según chi-cuadrado de Pearson con un gra-

do de libertad. Además, el análisis se completa con el estadístico Q_E, que permite comprobar si el modelo está bien especificado. Bajo la hipótesis nula de que el modelo está bien especificado, Q_E se distribuye según chi-cuadrado de Pearson con $k - 2$ grados de libertad.

Por último, es posible y aconsejable finalizar el análisis estadístico en un estudio de generalización de la fiabilidad proponiendo un modelo explicativo, basado en una regresión múltiple ponderada, que permita identificar el conjunto de variables moderadoras que mejor dan cuenta de la variabilidad de los coeficientes de fiabilidad. Variables tales como el número de ítems del test, la variabilidad de las puntuaciones del test, o la adaptación del test utilizada, entre otras muchas, es muy probable que configuren dicho modelo explicativo, ya que la teoría psicométrica así lo postula. De hecho, algunos autores proponen que en el análisis de las variables moderadoras se incorpore siempre la desviación típica de las puntuaciones del test para controlar su influjo sobre la variabilidad de los coeficientes de fiabilidad (cf. Rodríguez y Maeda, 2006).

Para la realización de los análisis estadísticos propios de un meta-análisis se puede utilizar cualquier paquete estadístico profesional (e.g., SAS, SPSS, STATA, SYSTAT, etc.). Pero son especialmente útiles los macros que algunos autores han elaborado para su ejecución en estos paquetes profesionales. Cabe mencionar a este respecto los macros elaborados por David B. Wilson para su uso en los programas SPSS, SAS y STATA.[6] También existen programas de software específicamente diseñados para realizar los análisis estadísticos propios de un meta-análisis, tales como *Comprehensive Meta-analysis 2.2* (Borenstein, Hedges, Higgins y Rothstein, 2005) o *MetaWin 2.0* (Rosenberg, Adams y Gurevitch, 1999).

Publicación

Un estudio de generalización de la fiabilidad finaliza con su publicación. La estructura del informe es similar a la de cualquier investigación: introducción, método, resultados y discusión (Botella y Gambara, 2002, 2006; Cooper, 1998; Rosenthal, 1995; Sánchez-Meca, 1999, 2003; Sánchez-Meca y Ato, 1989). En la introducción se presenta el test objeto de estudio, así como sus posibles versiones y campos de aplicación. En la sección de metodología lo habitual es presentar las secciones típicas de un estudio meta-analítico: (a) definición de los criterios de selección de los estudios empíricos para el meta-análisis; (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. Una minuciosa descripción de las decisiones tomadas a lo largo del estudio meta-analítico garantiza la máxima transparencia y la posibilidad de replicación del estudio por otros investigadores.

La sección de resultados debe comenzar con una descripción de las características de los estudios incluidos en el meta-análisis (poblaciones muestreadas, versiones del test, condiciones de aplicación, etc.). A continuación se suele presentar una estimación de la fiabilidad media para cada tipo de coeficiente de fiabilidad registrado, junto con una estimación por intervalo. El siguiente punto a tratar en los resultados es comprobar si existe heterogeneidad entre las estimaciones de la fiabilidad, es decir, si la fiabilidad es generalizable. Por regla general, los estudios presentarán una alta variabilidad en las estimaciones de la fiabilidad, por lo que un aspecto fundamental de los resultados implica analizar variables moderadoras que sean capaces de explicar esa variabilidad y contribuir, de esta forma, a entender los factores de los que depende la fiabilidad de las puntuaciones de un test. Los

[6] Pueden consultarse estos programas, e incluso obtenerlos libres de cargo, en la dirección web: http://mason.gmu.edu/~dwilsonb/ma.html. También puede acceder a la página web de la *Unidad de Meta-análisis* de la Universidad de Murcia para consultar una revisión del software existente actualmente sobre meta-análisis, en la dirección: http://www.um.es/facpsi/metaanalysis.

resultados deben ilustrarse, en la medida de lo posible, con tablas estadísticas y gráficos. En la sección de discusión y de conclusiones se deben relacionar los resultados obtenidos con los de otros posibles estudios de generalización de la fiabilidad similares, así como ofrecer una valoración del grado de fiabilidad que ofrecen las puntuaciones del test, su heterogeneidad y las variables que modelan tal heterogeneidad.

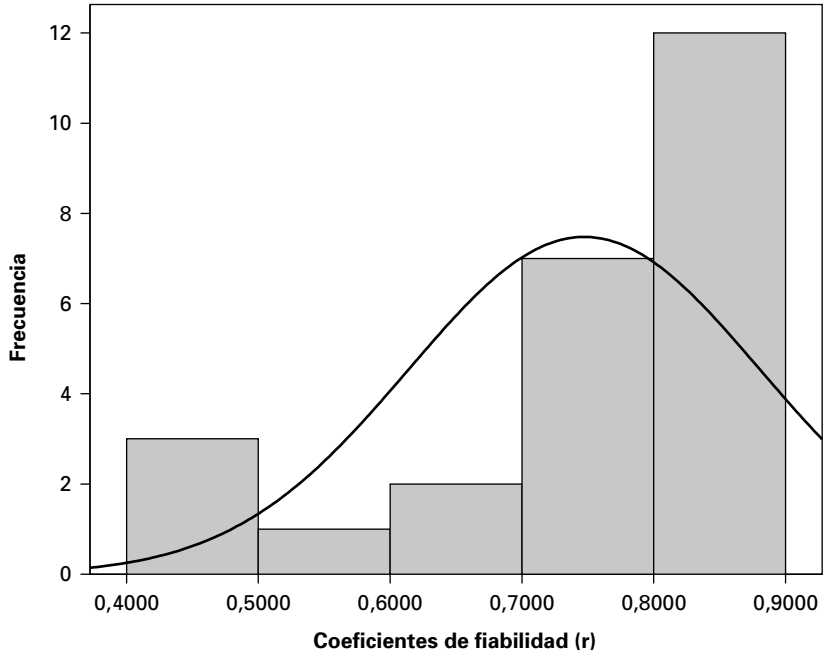
Por último, en la sección de referencias deben señalarse (por ejemplo, con un asterisco) los estudios meta-analizados y, si las limitaciones de espacio lo permiten, incluir un apéndice que recoja la base de datos completa con las principales variables analizadas.

Un ejemplo

Con objeto de ilustrar los cálculos necesarios para analizar estadísticamente los datos de un estudio de esta naturaleza, presentamos un ejemplo tomado de un estudio de generalización de la fiabilidad realizado por López-Pina, Sánchez-Meca y Rosa-Alcázar (2008) sobre la escala Hamilton de Depresión (*Hamilton Rating Scale for Depression*). El ejemplo también servirá para comprobar el grado de coincidencia en las estimaciones obtenidas con los diferentes modelos y métodos estadísticos que hemos presentado en el epígrafe anterior y que se están utilizando actualmente en este tipo de estudios.

En el Apéndice figura la base de datos utilizada para este ejemplo,[7] en el que se recogen 25 estudios que han aplicado dicha escala y reportaron una estimación de la fiabilidad mediante el cálculo del coeficiente alfa de Cronbach. Se observa una amplia variabilidad en los coeficientes de fiabilidad obtenidos, que va desde el valor más bajo, reportado en el Estudio 12, $r = 0.420$, hasta el más alto, reportado por el

Figura 1. Distribución de los coeficientes de fiabilidad (con el gráfico de la distribución normal superpuesto)



[7] En realidad, el estudio del test de Hamilton incluyó más muestras de las que aquí presentamos. Hemos preferido reducir el número de muestras para que el ejemplo sea más didáctico. En concreto, hemos seleccionado sólo aquellos coeficientes alfa que se obtuvieron a partir de la aplicación del test original de Hamilton, que contenía 17 ítems, dejando fuera modificaciones y adaptaciones del mismo.

Figura 2. Distribución de los coeficientes de fiabilidad transformados a *Z* de Fisher (con el gráfico de la distribución normal superpuesto)

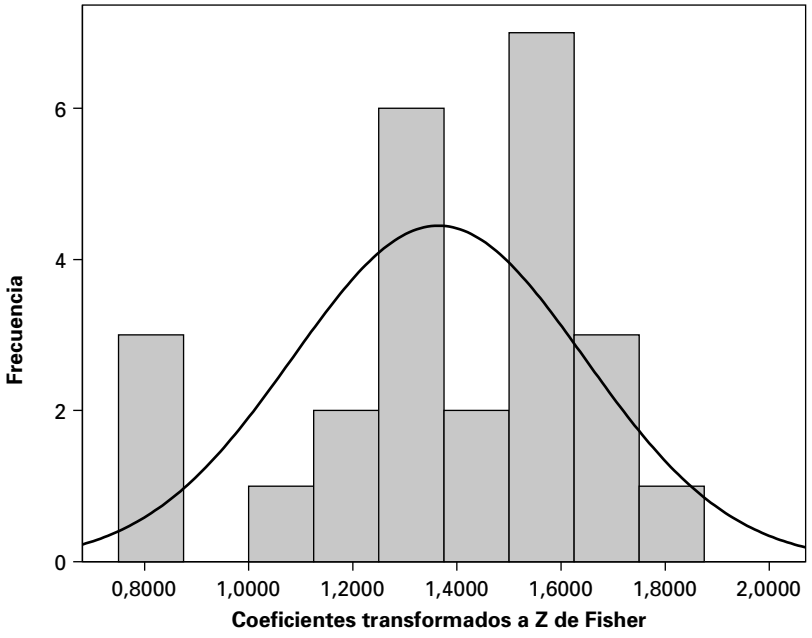
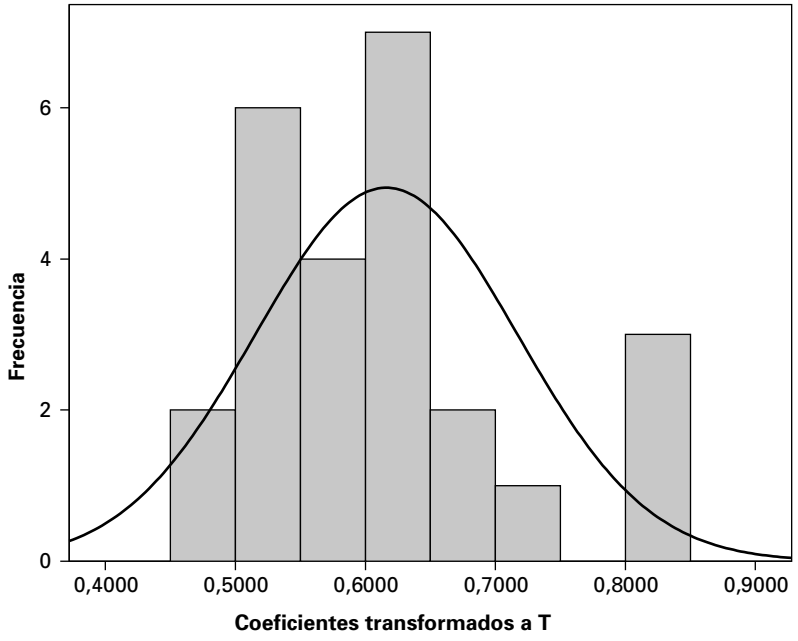


Figura 3. Distribución de los coeficientes de fiabilidad transformados a *T* según la ecuación (2) (con el gráfico de la distribución normal superpuesto)



Estudio 17, r = 0.897. En el Apéndice se incluye también la transformación a Z de Fisher de los coeficientes de fiabilidad, según la ecuación (1), y la transformación a puntuación T según la ecuación (2). Además, se incluye el tamaño muestral, N_i, y el número de ítems de cada aplicación del test, que en este caso fue constante (17 ítems). Por último, el Apéndice incluye dos variables moderadoras, una cualitativa y otra cuantitativa: el tipo de trastorno investigado en cada estudio (1: Depresión; 0: Otros trastornos) y la desviación estándar de las puntuaciones del test, S_i.

En la Figura 1 se presenta un histograma de los 25 coeficientes de fiabilidad incluidos en este meta-análisis. Una primera inspección de la distribución revela, como es de esperar en los coeficientes de fiabilidad, una marcada asimetría negativa (coeficiente de asimetría = -1.324; coeficiente de curtosis = 0.937), que da lugar a una clara desviación del supuesto de normalidad, como así lo evidencia el resultado estadísticamente significativo obtenido con la prueba de Shapiro-Wilk (p = .001). Las transformaciones a Z de Fisher y a puntuación T presentadas en las ecuaciones (1) y (2) tienen como objetivo normalizar la distribución de los coeficientes de fiabilidad. En efecto, como puede observarse en las Figuras 2 y 3, la distribución de las puntuaciones transformadas logra acercarse al supuesto de normalidad. En concreto, tanto la distribución de las puntuaciones Z (Figura 2) como la de las puntuaciones T (Figura 3) obtienen una menor asimetría que la exhibida por la distribución de los coeficientes de fiabilidad originales (puntuaciones Z: coef. asimetría = -0.707, coef. curtosis = -0.194; puntuaciones T: coef. asimetría = 0.821, coef. curtosis = -0.082), y la prueba de normalidad de Shapiro-Wilk no alcanzó la significación estadística en ninguno de los dos casos (p = .137 y p = .060, respectivamente), si bien las puntuaciones T rozaron la significación estadística.

Tabla 1. Coeficiente de fiabilidad medio con su intervalo de confianza al 95% obtenido por diferentes procedimientos de cálculo.

		I. C. al 95%			
Modelo estadístico	**Método**	r_+	r_i	r_s	**A.C.**
No ponderado: $w_i = 1$	r_i	0.747	0.695	0.799	0.104
	Z_i	0.877	0.849	0.900	0.051
	T_i	0.766	0.718	0.808	0.090
Ponderado por N_i: $w_i = N_i$	r_i	0.784	0.758	0.810	0.052
	Z_i	0.899	0.894	0.904	0.010
	T_i	0.805	0.777	0.830	0.053
Efectos Fijos: $w_i = \frac{1}{S^2_{r_i}} = \frac{1}{S^2_{Z_i}} = \frac{1}{S^2_{T_i}}$	r_i	0.852	0.844	0.859	0.015
	Z_i	0.899	0.895	0.905	0.010
	T_i	0.833	0.827	0.839	0.012
Efectos Aleatorios: $w_i^{EA} = \frac{1}{S^2_{r_i} + \tau^2} = \frac{1}{S^2_{Z_i} + \tau^2} = \frac{1}{S^2_{T_i} + \tau^2}$	r_i	0.767	0.735	0.799	0.064
	Z_i	0.877	0.845	0.903	0.058
	T_i	0.769	0.725	0.808	0.083

Los tres métodos representan el índice estadístico utilizado para los cálculos: los coeficientes de fiabilidad (r_i), la transformación a Z de Fisher de los coeficientes de fiabilidad (Z_i) y la transformación a puntuación T consistente en el cálculo de la raíz cúbica (T_i). A.C. es la amplitud confidencial del intervalo de confianza (I. C.). Y r_+ es el coeficiente de fiabilidad medio.

Por regla general, resulta conveniente transformar los coeficientes de fiabilidad originales para lograr una buena aproximación al supuesto de normalidad de la distribución. La elección entre uno u otro tipo de transformación debería tomarse combinando criterios teóricos y empíricos. En nuestro ejemplo, desde una perspectiva empírica, vemos que ambas transformaciones logran normalizar la distribución, aunque la transformación a *Z* de Fisher logra un mejor ajuste a la normalidad. Desde una perspectiva teórica, sería conceptualmente más apropiado utilizar la transformación *T*, ya que los coeficientes de fiabilidad meta-analizados son coeficientes alfa. Por tanto, ambas transformaciones tienen alguna razón de peso para ser seleccionadas. En lo que sigue, y con propósitos meramente didácticos, presentamos los cálculos meta-analíticos con los tres métodos: con los coeficientes de fiabilidad originales y con las dos transformaciones, *Z* y *T*.

El primer objetivo en un estudio de generalización de la fiabilidad es obtener un promedio de la fiabilidad exhibida por las diferentes aplicaciones del test, junto con una estimación por intervalo. En la Tabla 1 se recogen los resultados para los diferentes modelos estadísticos y métodos de estimación. Como puede observarse en dicha tabla, el coeficiente de fiabilidad medio varía según el procedimiento de cálculo utilizado, siendo el más bajo el obtenido directamente con los coeficientes de fiabilidad sin transformar y sin ponderar (r_+ = 0.747) y el más alto el obtenido mediante la transformación a *Z* de Fisher y asumiendo un modelo de efectos fijos, o bien ponderando por el tamaño muestral (r_+ = 0.899 en ambos casos).[8] Además, la amplitud confidencial del intervalo también varía sensiblemente, siendo el intervalo más ancho (menos preciso) el obtenido directamente con los coeficientes de fiabilidad y sin ponderar (A.C. = 0.104) y el más estrecho el obtenido con la transformación a *Z* de Fisher y mediante el modelo de efectos fijos o con ponderación por el tamaño muestral (A.C. = 0.010).

Tabla 2. Evaluación de la heterogeneidad exhibida por los coeficientes de fiabilidad según los tres métodos descritos: con los coeficientes de fiabilidad directamente (r_i), la transformación a *Z* de Fisher (Z_i) y la transformación a puntuación *T* (T_i).

Método	*Q*	*GL*	*p*	I^2	τ^2
r_i	311.97	24	< 0.0001	92.31%	0.0051
Z_i	495.90	24	< 0.0001	95.16%	0.0906
T_i	569.49	24	< 0.0001	95.78%	0.0078

Q: Prueba de heterogeneidad. *GL*: grados de libertad asociados a la prueba *Q* de heterogeneidad (en nuestro caso *GL* = *k* – 1 = 25 – 1 = 24). *p*: nivel crítico de probabilidad asociado a la prueba *Q*. I^2: índice de heterogeneidad. τ^2: estimación de la varianza inter-estudios.

El segundo objetivo suele ser evaluar la variabilidad de los coeficientes de fiabilidad con objeto de comprobar si ésta puede deberse al mero error de muestreo aleatorio o si, por el contrario, la variabilidad es tan grande que debe estar provocada por diversos factores (metodológicos, contextuales, poblacionales, etc.). En este último caso, tendremos que concluir que la fiabilidad de las puntuaciones del test no es generalizable a través de diferentes contextos,

[8] En realidad, el método de *Z* de Fisher ponderando por el tamaño muestral o asumiendo un modelo de efectos fijos son el mismo procedimiento, ya que el factor de ponderación es prácticamente el mismo: N_i y N_i – 3, respectivamente.

poblaciones y métodos de aplicación del test. La Tabla 2 recoge los resultados obtenidos con los tres métodos de cálculo (directamente con los coeficientes de fiabilidad, r_i, con la transformación a Z de Fisher, Z_i, o con la transformación a puntuación T, T_i). Cuando se analiza la variabilidad de los coeficientes de fiabilidad el único modelo estadístico apropiado es el modelo de efectos fijos. Como se puede observar en la Tabla 2, con los tres procedimientos de cálculo encontramos un estadístico de heterogeneidad, Q, estadísticamente significativo ($p < .0001$ en todos los casos), así como índices I^2 que se sitúan por encima del 90%. Por tanto, en este caso tenemos que concluir que la fiabilidad exhibida por las diferentes aplicaciones del test en cuestión no es generalizable.

Esta conclusión condiciona, al menos, dos aspectos. En primer lugar, que el modelo estadístico más plausible subyacente a los coeficientes de fiabilidad de este ejemplo es el modelo de efectos aleatorios y, en consecuencia, el cálculo del coeficiente de fiabilidad medio debe obtenerse asumiendo este modelo. En segundo lugar, que es preciso dar un tercer paso en el proceso de análisis estadístico consistente en buscar variables que sean capaces de explicar al menos parte de la heterogeneidad encontrada

Tabla 3. Resultados del ANOVA ponderado de efectos mixtos con los tres métodos de cálculo para la variable moderadora 'tipo de trastorno'.

		I. C. al 95%					
Método	**Trastorno**	r_+	r_i	r_s	Q_{wj}	*GL*	*p*
r_i	Depresión	0.788	0.752	0.825	37.978	14	.001
	Otros	0.730	0.677	0.782	21.218	9	.012
	Resultados ANOVA:	$Q_B(1) = 3.202$, $p = .073$ $Q_W(23) = 59.196$, $p < .0001$ $\omega^2 = 0.010$					
Z_i	Depresión	0.888	0.855	0.915	15.608	14	.338
	Otros	0.856	0.799	0.898	7.280	9	.608
	Resultados ANOVA:	$Q_B(1) = 1.349$, $p = .245$ $Q_W(23) = 22.889$, $p = .467$ $\omega^2 = 0.014$					
T_i	Depresión	0.789	0.741	0.830	21.892	14	.081
	Otros	0.735	0.661	0.796	11.005	9	.275
	Resultados ANOVA:	$Q_B(1) = 1.802$, $p = .179$ $Q_W(23) = 32.897$, $p = .083$ $\omega^2 = 0.010$					

Los tres métodos representan el índice estadístico utilizado para los cálculos: los coeficientes de fiabilidad (r_i), la transformación a Z de Fisher de los coeficientes de fiabilidad (Z_i) y la transformación a puntuación T consistente en el cálculo de la raíz cúbica (T_i). r_+ es el coeficiente de fiabilidad medio obtenido para cada categoría de la variable moderadora, que va acompañado de los límites confidenciales del intervalo de confianza al 95%. Q_{wj} es la prueba de homogeneidad intra-categoría. *GL* representa los grados de libertad de cada prueba Q_{wj}. p es el nivel de crítico de probabilidad asociado a cada prueba estadística. Q_B es la prueba de homogeneidad inter-categorías. Q_W es la prueba de homogeneidad global intra-categorías. ω^2 es el índice de proporción de varianza explicada 'omega cuadrada de Hays'.

entre los coeficientes de fiabilidad. Es en esta etapa en la que los coeficientes de fiabilidad actúan como la variable dependiente y las variables moderadoras como variables independientes o predictoras.

Respecto del primer aspecto comentado, el del cálculo del coeficiente de fiabilidad medio asumiendo el modelo de efectos aleatorios, en nuestro ejemplo lo encontramos en la Tabla 1, que alcanza los valores r_+ = 0.767, 0.877 y 0.769 para los tres métodos de cálculo, r_i, Z_i y T_i, respectivamente. Respecto del segundo, las Tablas 3 y 4 presentan sendos ejemplos de análisis de influjo de variables moderadoras de la variabilidad de los coeficientes de fiabilidad.

Como ejemplo de análisis de una variable moderadora cualitativa, en la Tabla 3 se presentan los resultados del ANOVA ponderado aplicado sobre cada método de cálculo (r_i, Z_i y T_i) y asumiendo un modelo de efectos mixtos.[9] La variable moderadora en cuestión fue el trastorno investigado en cada estudio, distinguiendo entre depresión y otros trastornos. En dicha tabla se puede comprobar cómo con los tres métodos de cálculo la fiabilidad media obtenida cuando el trastorno estudiado era la depresión fue mayor que cuando se estudiaban otros trastornos. Sin embargo, la prueba Q_B, que es la que evalúa si existen diferencias entre dichas medias, no alcanzó la significación estadística en ningún caso. Además, la proporción de varianza explicada por esta variable fue muy baja, estando en torno al 1% en los tres casos. Por tanto, la conclusión respecto del tipo de trastorno investigado es que no afectó a los coeficientes de fiabilidad obtenidos en las diferentes aplicaciones del test.

Tabla 4. Resultados del análisis de regresión ponderado de efectos mixtos para la variable moderadora 'desviación estándar de las puntuaciones del test'.

Método	*k*	*b*	*ET*	Q_R	*p*	Q_E	*p*	R^2
r_i	25	0.039	0.0086	20.749	< 0.0001	29.644	0.156	0.412
Z_i	25	0.089	0.0273	10.807	0.0010	13.824	0.932	0.439
T_i	25	-0.032	0.0091	12.743	0.0004	16.114	0.850	0.442

Los tres métodos representan el índice estadístico utilizado para los cálculos: los coeficientes de fiabilidad (r_i), la transformación a *Z* de Fisher de los coeficientes de fiabilidad (Z_i) y la transformación a puntuación *T* consistente en el cálculo de la raíz cúbica (T_i). *b* es el coeficiente de regresión no estandarizado asociado a la variable moderadora. *ET* es el error típico del coeficiente *b*. Q_R es la prueba de evaluación del modelo. Q_E es la prueba de especificación del modelo. *p* es el nivel de crítico de probabilidad asociado a cada prueba estadística. R^2 es el índice de proporción de varianza explicada.

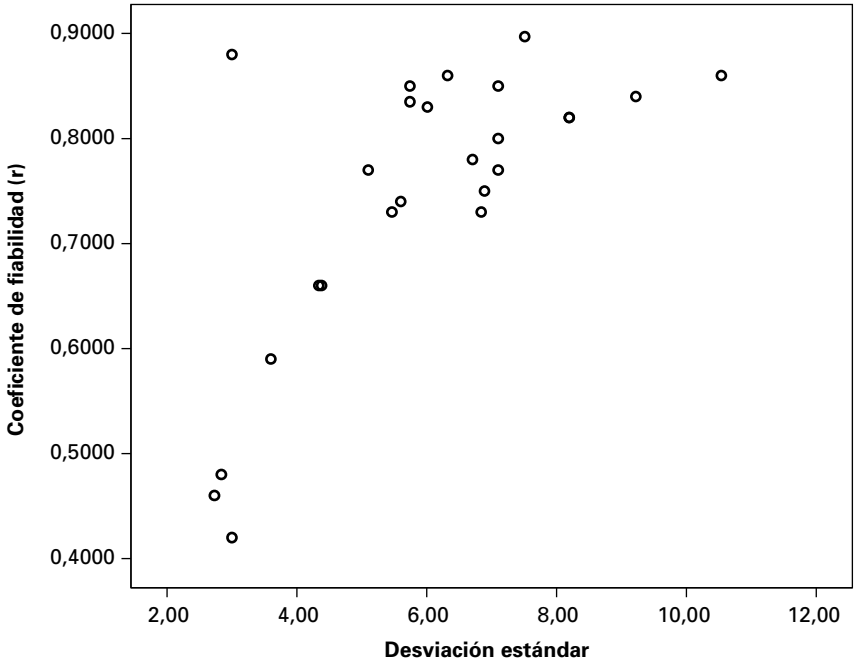
Para terminar con el ejemplo, presentamos en la Tabla 4 los resultados del análisis de regresión simple ponderado tomando como predictor la desviación estándar de las puntuaciones del test y asumiendo un modelo de efectos mixtos, para los tres métodos de cálculo. Como la teoría clásica de tests predice, la variabilidad de la muestra (representada aquí por la desviación estándar de las puntuaciones del test) está directamente relacionada con la fiabilidad. En efecto, los coeficientes de regresión asociados a la desviación estándar reflejan tal relación positiva con los coeficientes de fiabilidad,[10] siendo en los tres casos dicha relación estadísticamente

[9] Téngase en cuenta que, al haber encontrado en este ejemplo una alta heterogeneidad entre los coeficientes de fiabilidad, los modelos estadísticos más apropiados para examinar el influjo de variables moderadoras son los modelos de efectos mixtos.

[10] Téngase en cuenta que el signo negativo del coeficiente de regresión, *b* = -0.032, obtenido con las puntuaciones transformadas T_i se debe a que la transformación *T* definida en la ecuación (2) invierte el orden de los

significativa, según revela la prueba Q_R, con un porcentaje de varianza explicada que es superior al 40% en los tres métodos. Además, el modelo está bien especificado, también en los tres métodos, como así lo indica la ausencia de significación estadística de la prueba Q_E. Mediante un diagrama de dispersión, en la Figura 4 se ilustra la relación positiva encontrada entre los coeficientes de fiabilidad y las desviaciones estándar de las puntuaciones del test.

Figura 4. Diagrama de disapersión de la relación entre los coeficientes de fiabilidad y la desviación estándar de las puntuaciones del test.



Es preciso apuntar que hemos aplicado una multitud de métodos de cálculo diferentes sólo a efectos didácticos con el propósito de ilustrar cómo los resultados pueden variar dependiendo del método estadístico seguido. Pero debe quedar claro que la elección del procedimiento estadístico dependerá de diferentes aspectos y, muy en particular, del tipo de coeficiente de fiabilidad que se esté meta-analizando. En nuestro ejemplo, dado que los coeficientes de fiabilidad eran coeficientes alfa de Cronbach, y dado que obtuvimos una alta heterogeneidad entre éstos, el procedimiento estadístico más apropiado sería, en primer lugar, utilizar la transformación *T* para normalizar la distribución muestral de los coeficientes y, en segundo lugar, asumir un modelo de efectos aleatorios para estimar la fiabilidad media y modelos de efectos mixtos para examinar el influjo de variables moderadoras de tal variabilidad. Por tanto, en este ejemplo, la fiabilidad media más apropiada estaría representada por el valor $r_+ = 0.769$, con límites confidenciales 0.725 y 0.808 (ver Tabla 1). La prueba *Q* de heterogeneidad apropiada sería: $Q(24) = 569.49$, $p < .0001$, y con un índice $I^2 = 95.78\%$ (ver Tabla 2).

coeficientes originales, de forma que los coeficientes de fiabilidad más altos obtienen valores *T* más bajos que los coeficientes de fiabilidad bajos. Por tanto, ese signo negativo confirma la existencia de una relación positiva entre desviación estándar y fiabilidad.

Discusión y conclusiones

El propósito de este artículo fue presentar una panorámica de qué es el enfoque meta-analítico de generalización de la fiabilidad, 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. Finalmente, hemos ilustrado esta metodología con datos de un ejemplo real.

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 aplicados de que la fiabilidad es una propiedad del test, cuando realmente la fiabilidad 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 sobre esta muestra es 0.80».

En consecuencia, los investigadores en ciencias del comportamiento 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 (por ejemplo, a partir de la fiabilidad aportada en el manual del test). No cabe duda de que esta reciente metodología se encuentra todavía en fase de depuración y se requiere trabajo metodológico intenso para mejorar las técnicas estadísticas de integración. Cuestiones tales como si se deben transformar los coeficientes de fiabilidad y cuál es la transformación más apropiada, el problema de dependencia estadística cuando obtenemos varios coeficientes de fiabilidad en una misma muestra, el método idóneo de ponderación de los coeficientes o el modelo estadístico más apropiado para meta-analizar coeficientes de fiabilidad, no tienen todavía una respuesta definitiva. Otros problemas no abordados en este artículo, tales como el del sesgo de publicación o de no reporte de la fiabilidad, ya han recibido alguna atención (cf. el reciente artículo de Howell y Shields, 2008), pero necesitan también más investigación.

Pero aunque todavía queda mucho camino por recorrer para obtener una metodología depurada en los estudios de generalización de la fiabilidad, 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. Por último, no podemos olvidar el interés que los estudios de generalización de la fiabilidad tienen para la investigación teórica en psicometría, ya que nos permiten indagar en los factores metodológicos, contextuales, de sujeto y de procedimiento que pueden afectar a la fiabilidad de las puntuaciones de los tests.

Referencias

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.

Borenstein, M. J., Hedges, L. V., Higgins, J. C., & Rothstein, H. (2005). *Comprehensive meta-analysis: A computer program for research synthesis* (Vers. 2.2). Englewood, NJ: Biostat, Inc.

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

Botella, J. y Gambara, H. (2006). Doing and reporting a meta-analysis. *International Journal of Clinical and Health Psychology, 6*, 425-440.

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, 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.

Conway, J. M., Jako, R. A. y Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*, 565-579.

Cooper, H. M. (1998). *Integrating research: A guide for literature reviews* (2ª ed.). Thousand Oaks, CA: Sage.

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

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 15*, 297-334.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*, 481-489.

DerSimonian, R. y Laird, N. (1986). Meta-analysis of clinical trials. *Controlled Clinical Trials, 7*, 177-188.

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

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.

Glass, G. V., McGaw, B. y Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

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

Gulliksen, H. (1987). *Theory of mental tests*. Nueva York: Wiley.

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.

Harwell, M. (1997). An empirical study of Hedges's homogeneity test. *Psychological Methods, 2*, 219-231.

Hedges, L. V. (1994). Fixed effects models. En H. Cooper y L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285-299). Nueva York: Russell Sage Foundation.

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.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*, 177-189.

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*, 113-127.

Higgins, J. P. T. y Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539-1558.

Howell, R. T. y Shields, A. L. (2008). The file drawer problem in reliability generalization: A strategy to compute a fail-safe *N* with reliability coefficients. *Educational and Psychological Measurement, 68*, 120-128.

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

Hunter, J. E. y Schmidt, F. S. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.

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.

Hunter, J. E., Schmidt, F. S. y Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.

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.

Lipsey, M. W. y Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

López-Pina, J. A., Sánchez-Meca, J. y Rosa-Alcázar, A. I. (2008). *The Hamilton Rating Scale for Depression: A reliability generalization study*. Manuscrito en revisión, Universidad de Murcia.

Marín-Martínez, F., Sánchez-Meca, J., Huedo, T. y Fernández-Guzmán, I. (2007). Meta-análisis: ¿Dónde estamos y hacia dónde vamos? En A. Borges y P. Prieto (Eds.), *Psicología y ciencias afines en los albores del siglo XXI (Homenaje al profesor Alfonso Sánchez Bruno)* (pp. 87-102). Grupo Editorial Universitario.

Marín-Martínez, F. y Sánchez-Meca, J. (1998). Testing for dichotomous moderators in meta-analysis. *Journal of Experimental Education*, 67, 69-81.

Martín, J. L. R., Tobías, A. y Seoane, T. (Coords.) (2006). *Revisiones Sistemáticas en las Ciencias de la Vida*. Toledo: FISCAM.

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, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling*, *2*, 255-273.

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.

Parker, K. (1983). A meta-analysis of the reliability and validity of the Rorschach. *Journal of Personality Assessment, 47*, 227-231.

Parker, K. C. H., Hanson, R. K. y Hunsley, J. (1988). MMPI, Rorschach, and WAIS: A meta-analytic comparison of reliability, stability, and validity. *Psychological Bulletin, 103*, 367-373.

Pedhazur, E. J. y Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

Petrosino, A., Boruch, R. F., Soydan, H., Duggan, L. y Sánchez-Meca, J. (2001). Meeting the challenges of evidence-based policy: The Campbell Collaboration. *Annals of the American Academy of Political and Social Science*, *578*, 14-34.

Petticrew, M. y Roberts, H. (2006). *Systematic Reviews in the Social Sciences: A practical guide*. Malden, MA: Blackwell.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*, 173-184.

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

Rosenberg, M. S., Adams, D. C. y Gurevitch, J. (1999). *MetaWin: Statistical software for meta-analysis with resampling tests* (Vers. 2.0). Sunderland, MA: Sinauer Associates.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (2ª ed.). Newbury Park, CA: Sage.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118*, 183-192.

Rowley, G. L. (1976). The reliability of observational measures. *American Educational Research Journal, 13*, 51-59.

Salgado, J. F. y Moscoso, S. (1996). Meta-analysis of interrater reliability of job performance ratings in

validity studies of personnel selection. *Perceptual and Motor Skills, 83*, 1195-1201.

Sánchez-Meca, J. (1999). Meta-análisis para la investigación científica. En F. J. Sarabia (Coord.), *Metodología para la investigación en marketing y dirección de empresa* (pp. 173-201). Madrid: Pirámide.

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 (Eds.), *Tratado de psicología general I: Historia, teoría y método* (pp. 617-669). Madrid: Alhambra.

Sánchez-Meca, J., Boruch, R. F., Petrosino, A. y Rosa-Alcázar, A. I. (2002). La Colaboración Campbell y la práctica basada en la evidencia. *Papeles del Psicólogo*, 83, 44-48.

Sánchez-Meca, J. y Marín-Martínez, F. (1997). Homogeneity tests in meta-analysis: A Monte Carlo comparison of statistical power and Type I error. *Quality and Quantity, 31*, 385-399.

Sánchez-Meca, J. y Marín-Martínez, F. (1998). Testing continuous moderators in meta-analysis: A comparison of procedures. *British Journal of Mathematical & Statistical Psychology*, 51, 311-326.

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.

Schmidt, F. S. y Hunter, J. S. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529-540.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Hogrefe & Huber Pub.

Shadish, W. R., Chacón-Moscoso, S. y Sánchez-Meca, J. (2005). Evidence-based decision making: Enhancing systematic reviews of program evaluation results in Europe. *Evaluation*, *11*, 95-109.

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 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. y Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals. *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.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics, 30*, 261-293.

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine, 26*, 37-52.

Viswesvaran, 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.

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.

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

Yarnold, P. R. y Mueser, K. T. (1989). Meta-analyses of the reliability of Type A behaviour measures. *British Journal of Medical Psychology, 62*, 43-50.

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.

Apéndice. Base de datos del ejemplo ilustrativo tomado del estudio de generalización de fiabilidad de la *Escala Hamilton de Depresión* (López-Pina et al., 2008).

Apéndice. Base de datos del ejemplo ilustrativo tomado del estudio de generalización de fiabilidad de la *Escala Hamilton de Depresión* (López-Pina et al., 2008).

Estudio	N.º ítems	N_i	S_i	Trastorno	r_i	Z_i	T_i
1	17	112	5,46	0	0,7300	1,2722	0,6463
2	17	100	7,10	0	0,7700	1,3648	0,6127
3	17	50	4,38	0	0,6600	1,1341	0,6980
4	17	50	4,34	0	0,6600	1,1341	0,6980
5	17	41	6,71	0	0,7800	1,3900	0,6037
6	17	359	2,84	0	0,4800	0,8534	0,8041
7	17	97	3,60	0	0,5900	1,0157	0,7429
8	17	48	6,32	0	0,8600	1,6392	0,5192
9	17	89	5,74	0	0,8500	1,6019	0,5313
10	17	23	5,74	0	0,8350	1,5500	0,5485
11	17	94	6,89	1	0,7500	1,3170	0,6300
12	17	230	3,00	1	0,4200	0,7720	0,8340
13	17	165	5,60	1	0,7400	1,2942	0,6383
14	17	397	8,19	1	0,8200	1,5022	0,5646
15	17	472	10,54	1	0,8600	1,6392	0,5192
16	17	100	2,73	1	0,4600	0,8258	0,8143
17	17	921	7,51	1	0,8970	1,8029	0,4688
18	17	489	8,19	1	0,8200	1,5022	0,5646
19	17	135	5,10	1	0,7700	1,3648	0,6127
20	17	150	6,01	1	0,8300	1,5336	0,5540
21	17	120	6,84	1	0,7300	0,2722	0,6463
22	17	596	9,22	1	0,8400	1,5668	0,5429
23	17	289	7,10	1	0,8000	1,4436	0,5848
24	17	552	3,00	1	0,8800	1,7218	0,4932
25	17	49	7,10	1	0,8500	1,6019	0,5313

N_i: tamaño muestral. S_i: desviación estándar de las puntuaciones del test. Trastorno: Tipo de trastorno estudiado en la muestra (1: Depresión; =. Otros trastornos). r_i: coeficiente alpha de Cronbach. Z_i: transformación a *Z* de Fisher del coeficiente de fiabilidad mediante la Ecuación (1). T_i: transformación a puntuación *T* del coeficiente de fiabilidad mediante la Ecuación (2).