Psicothema 2012. Vol. 24, nº 1, pp. 161-166
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG


# Métodos para promediar coeficientes alfa en los estudios de generalización de la fiabilidad

José A. López-Pina, Julio Sánchez-Meca y José A. López-López
Universidad de Murcia

El enfoque de la generalización de la fiabilidad (GF) es un tipo de meta-análisis que pretende integrar un conjunto de coeficientes de fiabilidad obtenidos en varias aplicaciones de un test, con objeto de caracterizar el error de medida y determinar qué factores de los estudios pueden explicar su variabilidad. Se han propuesto en la literatura diferentes procedimientos para promediar un conjunto de coeficientes alfa independientes y no existe un consenso actual sobre qué métodos son los mejores. Presentamos los resultados de un estudio de simulación Monte Carlo para comparar el funcionamiento, en términos de sesgo y error cuadrático medio, de doce procedimientos propuestos por Feldt y Charter. Los procedimientos difieren en función de si los coeficientes se transforman o no, y de si se ponderan por el tamaño muestral o no. Nuestros resultados apuntan hacia la recomendación de que se utilicen procedimientos ponderados frente a los no ponderados, y a que se transformen los coeficientes mediante la propuesta de Hakstian y Whalen o la basada en la raíz cuadrada de la inversa del coeficiente alfa. Finalmente, se discuten las relaciones entre los diferentes procedimientos de promediar con los modelos estadísticos de efectos fijos, aleatorios y de coeficientes variables.

*Methods for averaging alpha coefficients in reliability generalization studies*. The reliability generalization (RG) approach is a kind of meta-analysis that aims to statistically integrate a set of independent reliability coefficients obtained in several applications of a test, with the purpose of characterizing the measurement error and determining which factors related to the studies and samples can explain its variability. Diverse procedures have been proposed in the literature for averaging a set of independent alpha coefficients, and there is no consensus about which methods are best. Here, we present the results of a Monte Carlo simulation study, comparing the performance of twelve procedures proposed in Feldt and Charter, in terms of bias and mean square error. These procedures differ from each other in the transformation (or not) of the coefficients, and in the application or not of a weighting scheme based on sample size. Our results recommend using weighted methods in contrast to unweighted ones, and transforming the coefficients by the Hakstian and Whalen's proposal or by the proposal based on the square root of the inverse alpha coefficient. Lastly, we discuss the relations between the diverse procedures for averaging alpha coefficients with fixed-effects, random-effects, and varying coefficients models.

Una práctica común en la investigación psicológica es inducir la fiabilidad de los tests que se utilizan en un estudio a partir de la reportada en estudios previos (generalmente, la obtenida en la validación o baremación original del test), en lugar de estimarla con los propios datos de la muestra (Vacha-Haase, Kogan y Thompson, 2000). Esta práctica se fundamenta en la creencia errónea de que la fiabilidad es una propiedad inherente al test, por lo que es habitual encontrar expresiones tales como «el coeficiente de fiabilidad del test es 0,80». Sin embargo, el coeficiente de fiabilidad cambia de una aplicación a otra en función de la variabilidad y composición de la muestra (Crocker y Algina, 1986; Suen, 1990). No es esperable obtener la misma fiabilidad cuando el test se aplica a población clínica, análoga o no clínica; o cuando se aplica a distintos grupos de edad, género, estatus sociocultural o contextos diferentes; o cuando el test se adapta a otras lenguas y/o culturas. De ahí que sea más apropiado hablar de la fiabilidad de las puntuaciones en una aplicación concreta del test que de la fiabilidad del test en sí mismo (Wilkinson y Task Force on Statistical Inference, 1999).

Dado que la fiabilidad de las puntuaciones puede variar de una aplicación a otra de un test, el meta-análisis se convierte en una metodología idónea para investigar cómo varía la fiabilidad en las diversas aplicaciones de un mismo test y si ésta puede generalizarse. En esta línea, Vacha-Haase (1998) propuso el enfoque de la generalización de la fiabilidad (GF) como un tipo de meta-análisis dirigido a examinar la varianza de error en los tests a través de sus diferentes aplicaciones e identificar las características de los estudios que podrían pronosticar esta variabilidad (Botella y Ponte, 2011; Botella y Suero, en prensa; Botella, Suero y Gambara, 2010; Henson y Thompson, 2002; Sánchez-Meca y López-Pina, 2008;

Fecha recepción: 7-6-11 • Fecha aceptación: 23-9-11
Correspondencia: José A. López-Pina
Facultad de Psicología
Universidad de Murcia
30100 Murcia (Spain)
e-mail: jlpina@um.es

Sánchez-Meca, López-Pina y López-López, 2008, 2009). Básicamente, en un estudio de GF se recogen los coeficientes de fiabilidad de los estudios que han aplicado un test y las variables que caracterizan a dichos estudios, de forma que los coeficientes de fiabilidad constituyen la variable dependiente, mientras que las características de los estudios actúan como variables predictoras de la variabilidad exhibida por los coeficientes de fiabilidad (Vacha-Haase, 1998; Vacha-Haase et al., 2000). En sus doce años de existencia ya se han realizado más de ochenta estudios de GF (Sánchez-Meca et al., 2008) con diferentes tests (e.g., Aguayo, Vargas, de la Fuente y Lozano, 2011; Kieffer y MacDonald, 2011; López-Pina, Sánchez-Meca y Rosa-Alcázar, 2009; Sánchez-Meca, López-Pina, López-López, Marín-Martínez, Rosa-Alcázar y Gómez Conesa, 2011).

La mayoría de los estudios de GF han integrado coeficientes alfa, ya que éste es el coeficiente de fiabilidad más comúnmente reportado en los estudios empíricos. Es por ello que el interés de este trabajo se centra en el coeficiente alfa. Si las varianzas de error son iguales en todos los ítems de un test, la fiabilidad de las puntuaciones de un test en una aplicación concreta del mismo se puede estimar mediante el coeficiente alfa, que es el promedio ponderado, en función de la longitud del test, entre las varianzas de los ítems y la varianza total del test (Cronbach, 1951).

Uno de los objetivos de un estudio de GF es obtener un promedio de los coeficientes de fiabilidad representativo de todos los estudios. Se han propuesto en la literatura diversos procedimientos para obtener dicho promedio (Feldt y Charter, 2006) y los estudios de GF han aplicado una gran variedad de ellos (Sánchez-Meca et al., 2008). Sin embargo, se han hecho pocos estudios de simulación para determinar qué método, o métodos, son los mejores. Los procedimientos más utilizados hasta ahora difieren en función de si se deben ponderar o no los coeficientes por el tamaño muestral y si se deben transformar previamente (e.g., mediante *Z* de Fisher).

Feldt y Charter (2006) presentaron seis procedimientos distintos para promediar coeficientes alfa, si bien éstos se convierten en 12 si tenemos en cuenta que cada uno de ellos puede aplicarse tanto ponderando como sin ponderar cada coeficiente por el tamaño muestral. Estos autores realizaron un estudio de simulación Monte Carlo para comparar su funcionamiento, para lo cual manipularon el número de estudios ($k$= 4, 10 y 50) y generaron, a partir de una población hipotética de coeficientes de fiabilidad, valores entre 0,70 y 0,95. Los resultados mostraron que las diferencias entre los seis procedimientos para calcular los promedios de los coeficientes alfa fueron tan pequeñas que no tienen relevancia práctica. Además, conforme aumentó el número de estudios, las diferencias promedio de los coeficientes de fiabilidad de los seis procedimientos aumentaron ligeramente, mientras que las desviaciones típicas se redujeron sensiblemente. Feldt y Charter (2006) concluyeron que no parece que haya una aproximación más correcta que las otras a la hora de obtener estos promedios, por lo que aconsejaron utilizar el primer procedimiento, consistente en promediar los coeficientes alfa sin transformar y sin ponderar, ya que su comprensión y estructura es más simple.

Aunque Feldt y Charter (2006) no encontraron diferencias apreciables entre los seis procedimientos, su estudio presenta varias limitaciones. En primer lugar, los coeficientes alfa se simularon a partir de un rango relativamente estrecho de valores (0,70-0,95). En segundo lugar, hicieron las simulaciones asumiendo que todos los estudios tenían el mismo tamaño muestral, lo cual es irreal en los estudios de GF. En tercer lugar, solo compararon el funcionamiento de los seis procedimientos sin ponderar, aunque los propios autores reconocían en su artículo que dichos procedimientos pueden también aplicarse ponderando por el tamaño muestral. En consecuencia, el propósito de este estudio fue comparar mediante simulación Monte Carlo, en términos del sesgo y el error cuadrático medio, el funcionamiento de los doce procedimientos bajo condiciones más realistas. Así, utilizamos un rango más amplio de coeficientes alfa y manipulamos el tamaño muestral de los estudios. De esta forma, pudimos comprobar si los métodos seguían presentando un comportamiento similar cuando se comparan bajo condiciones más parecidas a las que soportan los estudios de GF reales.

A continuación, se presenta una descripción de los doce métodos para promediar coeficientes alfa objeto de nuestro estudio comparativo, se aplican todos ellos a un ejemplo real y se presenta la metodología seguida en el estudio de simulación, sus resultados y las conclusiones.

La tabla 1 recoge los seis procedimientos presentados por Feldt y Charter (2006) tanto ponderados como no ponderados por el tamaño muestral. El primer procedimiento (P1) consiste en promediar los coeficientes alfa directamente sin transformarlos. En el procedimiento 2 (P2), Feldt y Charter (2006) definieron el coeficiente promedio como el valor que duplica la media de los errores típicos de medida. En el procedimiento 3 (P3) se asume que el coeficiente alfa es equivalente al coeficiente de fiabilidad obtenido por formas paralelas y que, como expresión de la correlación producto-momento de Pearson, podemos transformarlo a *Z* de Fisher, obtener el promedio ponderado y después retransformarlo a coeficiente alfa. En el procedimiento 4 (P4) se emplea la transformación de la raíz cúbica, $(1-\alpha)^{1/3}$, propuesta por Hakstian y Whalen (1976),

*Tabla 1*
Procedimientos para promediar coeficientes alfa

| Procedimiento | Promedio ponderado | Promedio no ponderado |
|---|---|---|
| P1 | $\mu_{1_P} = \frac{\sum n_j \hat{\alpha}_j}{\sum n_j}$ | $\mu_{1_{NP}} = \frac{\sum \hat{\alpha}_j}{k}$ |
| P2 | $\mu_{2_P} = 1 - \left[\frac{\sum n_j \left(1-\hat{\alpha}_j\right)^{1/2}}{\sum n_j}\right]^2$ | $\mu_{2_{NP}} = 1 - \left[\frac{\sum \left(1-\hat{\alpha}_j\right)^{1/2}}{k}\right]^2$ |
| P3 | $\mu_{3Z_P} = Tanghip\left[\frac{\sum \left(n_j-3\right) Z_{Alfa_j}}{\sum \left(n_j-3\right)}\right]$ | $\mu_{3Z_{NP}} = Tanghip\left[\frac{\sum Z_{Alfa_j}}{k}\right]$ |
| P4 | $\mu_{4P} = 1 - \left[\frac{\sum n_j \left(1-\hat{\alpha}_j\right)^{1/3}}{\sum n_j}\right]^3$ | $\mu_{4NP} = 1 - \left[\frac{\sum \left(1-\alpha_j\right)^{1/3}}{k}\right]^3$ |
| P5 | $\mu_{5_P} = \left(\frac{\sum n_j \sqrt{\hat{\alpha}_j}}{\sum n_j}\right)$ | $\mu_{5_{NP}} = \left(\frac{\sum \sqrt{\alpha_j}}{k}\right)^2$ |
| P6 | $\mu_{6Z_P} = \left[Tanhip\left(\frac{\sum \left(n_j-3\right) Z_{Indice_j}}{\sum \left(n_j-3\right)}\right)\right]^2$ | $\mu_{6Z_{NP}} = \left[Tanhip\left(\frac{\sum Z_{Indice_j}}{k}\right)\right]^2$ |

$\hat{\alpha}_j$: coeficiente alfa estimado en el estudio $j$. $n_j$: tamaño muestral del estudio $j$. $k$: número de coeficientes. $Z_{Alfa_j}$ y $Z_{Indice}$ son las transformaciones *Z* de Fisher para el coeficiente alfa y el índice de fiabilidad, respectivamente, del estudio $j$, según la ecuación general: $Z_j = \frac{1}{2} log \frac{1+\hat{\alpha}_j}{1-\hat{\alpha}_j}$. *Tanghip*: función trigonométrica 'tangente hiperbólica', que es la función inversa a la transformación *Z* de Fisher

que es una transformación normalizadora de la distribución de los coeficientes. En el procedimiento 5 (P5) se emplea el índice de fiabilidad, dado que éste es la raíz cuadrada del coeficiente de fiabilidad, con lo que se aproxima más a la estructura del coeficiente de correlación producto-momento de Pearson. Por último, el procedimiento 6 (P6) es equivalente al procedimiento P3 basado en la transformación *Z* de Fisher, pero con la diferencia de que el coeficiente de fiabilidad se sustituye por el índice de fiabilidad antes de ser transformado a *Z* de Fisher.

Para ilustrar el alcance de las diferencias que pueden obtenerse cuando se aplican los diferentes procedimientos para promediar coeficientes alfa, los hemos aplicado a los veinticinco coeficientes alfa recogidos en el estudio de GF realizado por López-Pina, Sánchez-Meca y Rosa-Alcázar (2009) sobre la escala Hamilton de Depresión. La tabla 2 presenta los promedios obtenidos con los doce procedimientos.

Se observa cierta discrepancia entre ellos, siendo los dos promedios más extremos 0,741 (obtenido con el procedimiento P5 no ponderado) y 0,810 (procedimiento P6 ponderado), lo que supone una discrepancia del 9,3% entre ambos. Se observa asimismo cómo, en este caso, sistemáticamente los promedios fueron más altos con los métodos ponderados que con los no ponderados. Ello se debe a la existencia de una ligera correlación, aunque no significativa, entre el coeficiente alfa y el tamaño muestral en este estudio de GF ($r$= 0,291, $p$= 0,158). Dentro de los procedimientos ponderados, los dos promedios más extremos fueron 0,777 (procedimiento P5) y 0,810 (procedimiento P6), lo que supone una discrepancia del 4,2% entre ambos. Entre los procedimientos no ponderados, los promedios más extremos fueron 0,741 (procedimiento P5) y 0,771 (procedimiento P3), con una discrepancia también en torno al 4%.

Las discrepancias encontradas entre los diferentes procedimientos en este ejemplo, aunque no son generalizables a otros estudios de GF, apuntan hacia la necesidad de examinar de forma sistemática sus propiedades estadísticas y su comparabilidad.

## Método

### *Generación de los coeficientes de fiabilidad*

En este estudio hemos partido de un coeficiente alfa paramétrico definido sobre una matriz de datos simulada formada por un millón de personas. Para asegurar la unidimensionalidad del test, utilizamos el modelo de Rasch unidimensional (Rasch, 1960/1980; Wright y Stone, 1979), ya que el coeficiente alfa solo está justificado utilizarlo cuando el test (o subtest) es esencialmente unidimensional (McDonald, 1999).

*Tabla 2*
Coeficientes alfa medios obtenidos con los 12 procedimientos aplicados al estudio de GF de López-Pina et al. (2009)

| Procedimiento | Ponderado | No ponderado |
|---|---|---|
| P1 | 0,784 | 0,747 |
| P2 | 0,800 | 0,762 |
| P3 | 0,810 | 0,771 |
| P4 | 0,805 | 0,766 |
| P5 | 0,777 | 0,741 |
| P6 | 0,810 | 0,770 |

Para obtener un coeficiente alfa paramétrico bajo se utilizó un test breve (10 ítems) cuyos parámetros de dificultad se distribuyeron uniformemente [-1, +1]. Para obtener un coeficiente alfa paramétrico elevado se utilizó un test largo (60 ítems) cuyos parámetros de dificultad se distribuyeron también uniformemente [-1, +1]. La distribución de la habilidad de las personas en ambas poblaciones fue normal [-3, +3]. El primer test produjo un coeficiente alfa paramétrico de 0,66410607, y el segundo de 0,93025883. Estos dos coeficientes de fiabilidad fueron tomados como los parámetros que permitieron evaluar el sesgo y el error cuadrático medio (ECM) de los promedios de los coeficientes de fiabilidad muestrales en las condiciones manipuladas en este estudio de simulación.

### *Condiciones experimentales*

Se emplearon tres tamaños muestrales promedio, $\bar{N}$, de 50, 100 y 150 sujetos, y el número de estudios se varió en los valores $k$= 10, 20, 30 y 40. Aunque el número de coeficientes alfa integrados en los estudios GF suele ser mayor del rango aquí utilizado (mediana= 51 estudios), decidimos mantener un rango de valores para esta variable próximo al utilizado por Feldt y Charter (2006) con propósitos de comparación. Ambas condiciones, $k$ y $\bar{N}$, se cruzaron completamente para cada uno de los dos coeficientes de fiabilidad paramétricos, lo que produjo un total de $2 \times 3 \times 4$= 24 combinaciones, cada una de las cuales fue replicada en 10.000 ocasiones.

En cada simulación de un estudio meta-analítico, el tamaño muestral de los estudios varió en función del tamaño muestral promedio diseñado. Así, para la condición $\bar{N}$= 50 se utilizaron los siguientes tamaños muestrales: 32, 36, 38, 40 y 104. Para $\bar{N}$= 100 se utilizaron los tamaños muestrales de: 64, 72, 76, 80 y 208; y para $\bar{N}$= 150 se utilizaron los tamaños muestrales de: 96, 108, 114, 120 y 312. Estos tamaños muestrales fueron repetidos sucesivamente en función del número de estudios incluidos en el meta-análisis. Por ejemplo, para $k$= 10 y $\bar{N}$= 50, los tamaños muestrales de los diez estudios fueron: 32, 32, 36, 36, 38, 38, 40, 40, 104 y 104. La distribución de los tamaños muestrales seleccionada es claramente asimétrica positiva, ya que ésa es la forma que adopta dicha distribución en los meta-análisis típicos del ámbito de la Psicología (Sánchez-Meca y Marín-Martínez, 1998).

### *Criterios de comparación*

Para evaluar el comportamiento de los doce procedimientos para promediar coeficientes alfa se emplearon dos criterios: el sesgo y el error cuadrático medio (ECM). El ECM se calculó mediante

$$ECM = \frac{\sum\left(\bar{\hat{\alpha}}_j - \alpha\right)^2}{10.000} \quad ,$$

donde $\bar{\hat{\alpha}}_j$ es el promedio de los coeficientes de fiabilidad, ponderados o no ponderados, en cada uno de los seis procedimientos, y $\alpha$ es el coeficiente de fiabilidad paramétrico. Este criterio permite evaluar el grado de variabilidad que exhibe cada procedimiento en la estimación del coeficiente alfa paramétrico. Debe tenerse en cuenta que el *ECM* aquí definido es un indicador que está en función de dos factores: de la variabilidad del estimador en torno a su propia media, $Var(\bar{\hat{\alpha}}_j)$, y del sesgo de dicho estimador, según la ecuación; $ECM = Var(\bar{\hat{\alpha}}_j) + Sesgo^2$. El *ECM* nos informa, pues, de

la estabilidad del estimador. El sesgo de cada procedimiento en la estimación del coeficiente alfa paramétrico se obtuvo mediante

$$Sesgo = \frac{\sum \hat{\bar{\alpha}}_j}{10.000} - \alpha_{paramétrico}$$

## Resultados

### *Error cuadrático medio*

La tabla 3 presenta los promedios de los ECMs para los doce procedimientos. Téngase en cuenta que todas las tablas de resultados (tablas 3-6) presentan los datos multiplicados por 10.000 para facilitar su interpretación. Las estimaciones no ponderadas obtuvieron ECMs mayores que los promedios ponderados. También se aprecian diferencias consistentes según el valor del coeficiente alfa paramétrico. Cuando el coeficiente alfa poblacional fue elevado ($\alpha$= 0,93), las diferencias en los ECMs entre los doce procedimientos exhibieron variaciones despreciables, entre 0,012 y 0,016. Cuando el coeficiente alfa paramétrico estuvo en torno a la fiabilidad esperada en un estudio experimental ($\alpha$= 0,66), tampoco se encontraron diferencias apreciables entre los seis procedimientos cuando se ponderó por el tamaño muestral, pero si no se pondera, P1 fue el que obtuvo un ECM más elevado (Media= 61,208) en comparación con el resto de procedimientos, que variaron entre 1,123, para P4, y 1,297, para P5.

La tabla 4 presenta el ECM de cada procedimiento del coeficiente alfa en función del tamaño muestral medio de los estudios ($\bar{N}$). En todos los procedimientos el ECM disminuyó con el aumento del tamaño muestral. Además, la ponderación de los coeficientes alfa produjo una disminución apreciable del ECM sobre la no ponderación. En general, los procedimientos P2 y P4 fueron los que exhibieron menores ECMs, mientras que los procedimientos P1 y P5 fueron los que mostraron mayores ECMs. Resultados similares se observaron en función del número de estudios, por lo que se obvia la presentación de la tabla de resultados: para valores mayores en el número de estudios se observó una reducción del ECM en todos los procedimientos, ponderados y no ponderados, y mejores resultados para P2 y P4 y peores para P1 y P5.

*Tabla 3*
Media (y desviación típica) del ECM de los 12 procedimientos para promediar coeficientes alfa (multiplicadas por 10.000)

| Procedimiento | $\alpha$= 0,66 | | $\alpha$= 0,93 | |
|---|---|---|---|---|
| | No ponderado | Ponderado | No ponderado | Ponderado |
| P1 | 61,208 (2,541) | 1,004 (2,045) | 0,015 (0,028) | 0,012 (0,024) |
| P2 | 1,138 (2,312) | 0,957 (1,908) | 0,016 (0,029) | 0,013 (0,024) |
| P3 | 1,128 (2,258) | 0,956 (1,875) | 0,016 (0,031) | 0,014 (0,025) |
| P4 | 1,123 (2,264) | 0,951 (1,880) | 0,016 (0,030) | 0,013 (0,025) |
| P5 | 1,297 (2,788) | 1,060 (2,190) | 0,015 (0,028) | 0,012 (0,024) |
| P6 | 1,134 (2,285) | 0,958 (1,890) | 0,016 (0,031) | 0,014 (0,025) |

### *Sesgo de los estimadores*

Como se puede observar en la tabla 5, todos los procedimientos dieron coeficientes alfa promedio sesgados positivamente para $\alpha$= 0,93. Por otra parte, para $\alpha$= 0,66 se apreciaron también algunas diferencias entre los distintos procedimientos, aunque no en el mismo sentido que en el caso anterior. Así, los procedimientos P3, P4 y P6 sobreestimaron el coeficiente alfa paramétrico, mientras que los procedimientos P1 y P5 lo infraestimaron. El procedimiento que obtuvo promedios menos sesgados fue P2, tanto ponderado (Media= 0,908) como no ponderado (Media= -0,570), seguido por P4 (Media= 4,537 y 5,224), P6 (Media= 6,692 y 7,056) y P3 (Media= 8,401 y 8,485). Muy alejados de estos valores estuvieron P1 (Media= -15,990 y -12,146) y P5 (Media= -24,434 y -19,245). Siguiendo con el caso de $\alpha$= 0,66, los procedimientos menos sesgados fueron de nuevo P2 (Media= -0,570 y 0,908) y P4 (Media= 4,537 y 5,224), mientras que los más sesgados fueron, nuevamente, P1 (Media= -15,990 y -12,146) y P5 (Media= -24,434 y -19,245).

*Tabla 4*
Media (y desviación típica) del ECM en función del tamaño muestral (multiplicadas por 10.000)

| Procedimiento | $\alpha$= 0,66 | | | | | |
|---|---|---|---|---|---|---|
| | No ponderado | | | Ponderado | | |
| | 50 | 100 | 150 | 50 | 100 | 150 |
| P1 | 2,135 (3,856) | 0,878 (1,424) | 0,612 (1,074) | 1,704 (3,026) | 0,786 (1,350) | 0,520 (0,893) |
| P2 | 1,944 (3,468) | 0,851 (1,325) | 0,604 (1,049) | 1,585 (2,793) | 0,769 (1,312) | 0,516 (0,879) |
| P3 | 1,923 (3,369) | 0,854 (1,369) | 0,607 (1,045) | 1,575 (2,732) | 0,772 (1,308) | 0,521 (0,879) |
| P4 | 1,916 (3,384) | 0,850 (1,367) | 0,604 (1,044) | 1,569 (2,743) | 0,768 (1,306) | 0,517 (0,879) |
| P5 | 2,352 (4,268) | 0,912 (1,476) | 0,625 (1,101) | 1,840 (3,268) | 0,810 (1,390) | 0,528 (0,909) |
| P6 | 1,939 (3,418) | 0,655 (1,373) | 0,607 (1,047) | 1,582 (2,758) | 0,773 (1,311) | 0,520 (0,880) |

*Tabla 5*
Media (y desviación típica) del sesgo para los 12 procedimientos propuestos para promediar coeficientes alfa (multiplicadas por 10.000)

| Procedimiento | $\alpha$= 0,66 | | $\alpha$= 0,93 | |
|---|---|---|---|---|
| | No ponderado | Ponderado | No ponderado | Ponderado |
| P1 | -15,990 (108,755) | -12,146 (99,438) | 3,865 (11,506) | 3,307 (10,606) |
| P2 | -0,570 (106,432) | 0,908 (97,800) | 4,709 (11,549) | 4,023 (10,635) |
| P3 | 8,401 (105,885) | 8,485 (97,399) | 5,487 (11,614) | 4,629 (10,673) |
| P4 | 4,537 (105,882) | 5,224 (97,405) | 4,990 (11,569) | 4,262 (10,649) |
| P5 | -24,434 (111,214) | -19,245 (101,128) | 3,802 (11,505) | 3,253 (10,606) |
| P6 | 6,692 (106,266) | 7,056 (97,641) | 5,485 (11,614) | 4,624 (10,673) |

La tabla 6 presenta el sesgo de los seis procedimientos en función del tamaño muestral para $\alpha$= 0,66. Los procedimientos P1 y P5 produjeron infraestimaciones del parámetro del coeficiente de fiabilidad cuando se ponderaron los coeficientes de fiabilidad, mientras que el procedimiento P2 solo infraestimó el parámetro con un tamaño muestral bajo. El resto de procedimientos (P3, P4 y P6) sobreestimaron el coeficiente de fiabilidad paramétrico independientemente del tamaño muestral promedio. Además, tomando el sesgo en términos absolutos, observamos que el sesgo de las estimaciones de los promedios ponderados se redujo en los procedimientos P1 y P5 en función del tamaño muestral, mientras que en el resto de procedimientos tuvo un comportamiento algo menos estable. Resultados similares se obtuvieron en función del número de estudios, por lo que se obvia la presentación de la tabla.

*Tabla 6*
Media (y desviación típica) del sesgo en función del tamaño muestral (multiplicadas por 10.000)

| Procedimiento | $\alpha$= 0,66 | | | | | |
|---|---|---|---|---|---|---|
| | No ponderado | | | Ponderado | | |
| | 50 | 100 | 150 | 50 | 100 | 150 |
| P1 | -33,283 (142,260) | -10,488 (93,122) | -4,199 (78,137) | -26,077 (127,902) | -7,954 (88,304) | -2,406 (72,095) |
| P2 | -7,174 (139,228) | 1,675 (92,242) | 3,790 (77,610) | -4,409 (125,807) | 2,585 (87,635) | 4,547 (71,718) |
| P3 | 7,871 (138,458) | 8,821 (91,997) | 8,512 (77,465) | 8,075 (125,228) | 8,748 (87,447) | 8,632 (71,629) |
| P4 | 1,430 (138,400) | 5,708 (91,999) | 6,444 (77,462) | 2,736 (125,221) | 6,080 (87,439) | 6,857 (71,610) |
| P5 | -48,011 (145,669) | -16,934 (93,989) | -8,358 (78,631) | -38,195 (130,178) | -13,522 (88,994) | -6,018 (72,444) |
| P6 | 4,876 (139,150) | 7,525 (92,168) | 7,674 (77,558) | 5,646 (125,668) | 7,626 (87,570) | 7,895 (71,677) |

## Discusión y conclusiones

En el ámbito de los estudios de GF, Feldt y Charter (2006) propusieron diferentes métodos para promediar un conjunto de coeficientes alfa independientes obtenidos en diferentes aplicaciones de un mismo test. Con objeto de extender los resultados de su estudio de simulación, hemos realizado una simulación Monte Carlo bajo condiciones más amplias y realistas en términos de distribución de los tamaños muestrales de los estudios meta-analizados y del número de estudios. Nuestros resultados no coinciden plenamente con los obtenidos por Feldt y Charter (2006) ya que, a diferencia de lo que ellos concluyeron, nuestros datos apuntan hacia la existencia de diferencias entre los métodos de promediar coeficientes alfa. En concreto, los dos procedimientos que muestran los mejores resultados, en general, en términos de ECM y sesgo del estimador, son P2 y P4, que consisten en aplicar la transformación $(1-\hat{\alpha}_j)^{1/2}$ y la transformación de Hakstian y Whalen (1976), respectivamente. Además, se tiende a obtener mejores resultados cuando se pondera en función del tamaño muestral que cuando no se ponderan los coeficientes. Por el contrario, los procedimientos con peores resultados fueron P1 y P5, que consisten en calcular el promedio con los propios coeficientes de fiabilidad sin transformar y con el índice de fiabilidad, respectivamente.

Nuestros resultados tienen varias implicaciones para la práctica de los estudios de GF. En primer lugar, el hecho de que se obtengan mejores resultados, en términos de sesgo y ECM, con los métodos ponderados coincide con la recomendación de Henson y Thompson (2002). En segundo lugar, el hecho de que los métodos basados en calcular directamente los coeficientes alfa sin transformar (P1) ofrezcan resultados poco satisfactorios va en contra de la recomendación recientemente planteada por Bonett (2010), según la cual el promedio de coeficientes alfa debería hacerse sin aplicar ninguna transformación de éstos. Bajo las condiciones manipuladas en nuestro estudio Monte Carlo, nuestros resultados desaconsejan esta práctica. Posiblemente, el carácter asimétrico de la distribución de los coeficientes alfa apunta a la conveniencia de normalizarla mediante alguna transformación.

En nuestro estudio comparativo no hemos planteado *a priori* ningún modelo estadístico de partida. Sin embargo, es habitual asumir alguno cuando se realiza un estudio de GF. De los diferentes modelos estadísticos propuestos en la literatura meta-analítica, los más reconocidos actualmente son los modelos de efectos fijos y de efectos aleatorios (Borenstein, Hedges, Higgins y Rothstein, 2009, 2010) y, muy recientemente, el modelo de coeficientes variables propuesto por Bonett (2010). Este último coincide con el procedimiento P1 sin ponderar, y el grado de generalización de los resultados solo puede extenderse a una población de estudios idéntica a los incluidos en el estudio de GF. Este grado de generalización es el mismo que se asume desde el modelo de efectos fijos, si bien bajo este modelo el procedimiento para promediar los coeficientes implica aplicar alguna transformación normalizadora de su distribución y, además, ponderar los coeficientes en función de la inversa de la varianza de éstos. En sentido estricto, ninguno de los doce procedimientos propuestos por Feldt y Charter (2006) se ajusta a este modelo, por lo que sería conveniente ampliar los resultados de nuestra simulación en el futuro. En cuanto al modelo de efectos aleatorios, los procedimientos basados en la aplicación de alguna transformación normalizadora de la distribución y que ponderan por el tamaño muestral pueden asimilarse a este modelo, si bien es más recomendable ponderar por la inversa de la varianza de los coeficientes, definida ésta como la suma de dos varianzas: la varianza intra-estudio y la varianza inter-estudios (Borenstein et al., 2009). Si el meta-analista desea generalizar sus resultados a una población mayor de estudios no exactamente idénticos a los incluidos en el meta-análisis, entonces el modelo de elección debería ser el de efectos aleatorios.

Los resultados de nuestro estudio deberían, pues, tomarse como una primera aproximación al problema de cómo promediar coeficientes alfa y, como tal, futuros estudios deberían incluir otros procedimientos propuestos, tales como métodos más sofisticados de ponderación, arriba comentados, así como la transformación de Bonett (2002), que Feldt y Charter (2006) no recogieron y que, sin embargo, fue propuesta como una transformación normalizadora de la distribución del coeficiente alfa y estabilizadora de su varianza.

Finalmente, se hace preciso algún comentario sobre el impacto real que en estudios de GF puede tener el hecho de utilizar diferentes procedimientos para promediar coeficientes alfa. Debe tenerse en cuenta que las diferencias encontradas entre los procedimientos las hemos discutido (y presentado en las tablas) multiplicadas por 10.000, para facilitar su interpretación. Pero ello conlleva la duda de si tales diferencias pueden no tener implicaciones tangibles en los estudios de GF reales. Atendiendo al ejemplo ilustrado en la tabla 2, consideramos que tales diferencias pueden conllevar

cambios relevantes en el promedio obtenido. Estas apreciaciones coinciden con otras comparaciones que hemos realizado en otro lugar con datos reales de varios estudios de GF (Sánchez-Meca, López-López y López-Pina, 2011).

Agradecimiento

Proyecto financiado por la Fundación Séneca (08650/PHCS/08) de la comunidad autónoma de Murcia.

## Referencias

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

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

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

Borenstein, M., Hedges, L.V., Higgins, J.P.T., y Rothstein, H.R. (2010). A basic introduction to fixed-effect and random-effects models in meta-analysis. *Research Synthesis Methods, 1*, 97-111.

Botella, J., y Ponte, G. (2011). Effects of the heterogeneity of the variances on reliability generalization: An example with the Beck Depression Inventory. *Psicothema, 23*, 516-522.

Botella, J., y Suero, M. (en prensa). Managing heterogeneity of variances in studies of internal consistency generalization. *Methodology*. DOI: 10.1027/1614-2241/a000039.

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

Crocker, L., y Algina, J. (1986). *An 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.

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

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

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

Kieffer, K.M., y MacDonald, G. (2011). Exploring factors that affect score reliability and validity in the Ways of Coping Questionnaire reliability coefficients: A meta-analytic reliability generalization study. *Journal of Individual Differences, 32*, 26-38.

López-Pina, J.A., Sánchez-Meca, J., y Rosa-Alcázar, A.I. (2009). The Hamilton Rating Scale for Depression: A meta-analytic reliability generalization study. *International Journal of Clinical and Health Psychology, 9*, 143-159.

McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: LEA.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks paedogoiske Institut (Chicago, IL: University Chicago Press, 1980).

Sánchez-Meca, J., López-López, J.A, y López-Pina, J.A. (2011). *Some recommended practices when reliability generalization (RG) studies are conducted*. Manuscrito no publicado, Universidad de Murcia.

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

Sánchez-Meca, J., López-Pina, J.A., y López-López, J.A. (2008). Una revisión de los estudios meta-analíticos de generalización de la fiabilidad. *Escritos de Psicología*, *1-2*, 107-118.

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

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

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

Suen, H.K. (1990). *Principles of test theories*. Hillsdale, NJ: LEA.

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., 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.

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

Wright, B.D., y Stone, M. (1979). *Best test design*. Chicago, IL: MESA Press.