Recibido / Recebido: 20.11.2023 - Aceptado / Aceite: 18.01.2024 https://doi.org/10.21865/RIDEP72.2.11

Comparación de Métodos Estadísticos de Meta-Análisis de Generalización de la Fiabilidad con Datos Reales

A Comparison of Statistical Methods of Reliability Generalization Meta-Analysis with Real Datasets

Raimundo Aguayo-Estremera[1]

Resumen

La fiabilidad es una propiedad que deben mostrar todos los tests. Dado que no es extrapolable de una muestra a otra, para estudiarla se desarrolló el meta-análisis de generalización de la fiabilidad. Actualmente existen muchos métodos estadísticos, pero no hay acuerdo sobre el más adecuado para este tipo de meta-análisis. Este trabajo compara en tres conjuntos de datos reales 15 métodos estadísticos distintos en relación a la distribución de coeficientes de fiabilidad, la estimación de la fiabilidad promedio y el efecto de las variables moderadoras. Los resultados indicaron que las distribuciones tienden a normalizarse con las transformaciones. Las discrepancias en las estimaciones de la fiabilidad promedio fueron pequeñas, pero pueden llevar a distintas conclusiones sustantivas. Las variables moderadoras obtuvieron resultados significativos en función del método empleado. Se debate la forma en que esta discrepancia entre métodos puede repercutir en aplicaciones concretas de estudios de generalización de la fiabilidad.

Palabras clave: generalización de la fiabilidad, meta-análisis, métodos estadísticos, alfa, psicometría

Abstract

The reliability is a property that all tests must exhibit. Since it is not generalizable from one sample to another, the meta-analysis of reliability generalization was developed. Currently, there are many statistical methods, but there is no agreement on the most suitable for this type of meta-analysis. This study compares 15 different statistical methods in three real datasets regarding the distribution of reliability coefficients, the estimation of average reliability, and the effect of moderator variables. The results indicated that distributions tend to achieved normality with the transformations. Discrepancies in estimates of average reliability were small but could lead to different substantive conclusions. Moderator variables yielded significant results depending on the method employed. The discussion revolves around how this discrepancy between methods may influence on specific applications of reliability generalization studies.

Keywords: reliability generalization, meta-analysis, statistical methods, alpha, psychometrics

[1]Doctor en Psicología. Profesor en el Departamento de Psicobiología y Metodología de las Ciencias del Comportamiento. Facultad de Psicología (Campus de Somosaguas), Universidad Complutense de Madrid. Carretera de Húmera s/n, 28223 Pozuelo de Alarcón, Madrid, España. E-mail: raaguayo@ucm.es.

ISSN: 1135-3848 print /2183-6051online

Introducción

La fiabilidad es una propiedad psicométrica muy importante al aplicar un test a una muestra de sujetos, que está relacionada con la precisión de las medidas y tiene consecuencias relevantes en la investigación aplicada. Por ejemplo, afecta a la potencia y a la precisión de los métodos estadísticos inferenciales, impacta en el sesgo de los estimadores de la pendiente en un análisis de regresión y determina el grado de atenuación que se produce en la diferencia de medias estandarizada y en la correlación bivariada (Bonett, 2010). La fiabilidad es una propiedad de las puntuaciones obtenidas en una administración específica con una muestra concreta, no del propio test. Por lo tanto, puede variar entre muestras si también varían la composición, variabilidad y condiciones de administración (Crocker & Algina, 1986; Grondlund & Linn, 1990; Romano & Kromrey, 2009). Así, se puede esperar que la fiabilidad cambie al aplicar el mismo test a diferentes grupos de población (por ejemplo, hombres/mujeres, clínicos/no clínicos) o en contextos diferentes (por ejemplo, países, culturas, idiomas).

Dado que no se puede asumir que la fiabilidad de un test concreto sea generalizable a lo largo de distintas administraciones, una forma útil de estudiarla consiste en la integración cuantitativa de los coeficientes de fiabilidad obtenidos en distintos estudios empíricos a partir del meta-análisis. Vacha-Haase (1998) introdujo el enfoque meta-analítico de generalización de la fiabilidad (GF), que consiste en un meta-análisis en el que se integran los coeficientes de fiabilidad obtenidos en diferentes aplicaciones de un test. Los tres objetivos principales de los estudios de GF son: (a) calcular una estimación promedio de fiabilidad; (b) evaluar la cantidad de variabilidad en los coeficientes de fiabilidad; y (c) examinar las fuentes de variabilidad entorno a la fiabilidad promedio (Botella et al., 2010).

Desde el trabajo inicial de Vacha-Haase (1998), se han desarrollado numerosos métodos estadísticos para llevar a cabo un estudio GF, y actualmente no existe acuerdo sobre qué métodos deben utilizarse. Por lo tanto, los investigadores tienen la libertad de elegir entre distintos métodos estadísticos (Botella & Ponte, 2011; Feldt & Charter, 2006; Henson & Thompson, 2002; Mason et al., 2007; Rodríguez & Maeda, 2006; Rouse, 2007; Vacha-Haase & Thompson, 2011). Curiosamente, esta libertad no ha sido considerada como problemática por los precursores del enfoque de GF. Sin embargo, esta diversidad de métodos debe tratarse con precaución, ya que se pueden obtener conclusiones diferentes según el método estadístico seleccionado (Sánchez-Meca et al., 2013) y los investigadores podrían incurrir en prácticas de investigación cuestionables (Paniagua et al., 2022), como cambiar de método tras haber observado los resultados del análisis de moderadores.

Los métodos estadísticos en el enfoque GF difieren fundamentalmente en tres aspectos (Sánchez-Meca & López-Pina, 2008): el modelo estadístico subyacente (modelos de efecto fijo o de efectos aleatorios), el factor de ponderación de los coeficientes de fiabilidad (por ejemplo, el inverso de la varianza o el tamaño muestral) y el método de transformación de los coeficientes de fiabilidad (por ejemplo, la transformación z de Fisher). En la Tabla 1 se puede ver un resumen de los procedimientos estadísticos analizados en este trabajo.

Los dos primeros problemas se han abordado conjuntamente a partir del desarrollo de los distintos modelos estadísticos. Se han propuesto seis modelos en la literatura: (a) mínimos cuadrados ordinarios (MCO) (Vacha-Haase, 1998); (b) de efecto fijo (EF) (Hedges & Olkin, 1985; Konstantopoulos & Hedges, 2009); (c) de coeficientes variables (CV), desarrollado por Laird y Mosteller (1990) y respaldado por Bonett (2008, 2009, 2010); (d) de efectos aleatorios (RE) (Hedges & Vevea, 1998; Raudenbush, 2009); (e) de efectos aleatorios corregido (RE-c) (Hartung, 1999; Knapp & Hartung, 2003); y (f) de efectos aleatorios ponderando por el tamaño muestral (RE-n), que ha sido desarrollado por Hunter y Schmidt (2004). Las consecuencias de asumir un modelo estadístico dado afectan la medida en que los resultados pueden generalizarse: los modelos EF y VC buscan generalizar los resultados solo a estudios con características similares a los incluidos en el meta-análisis, mientras que los modelos EA están pensados para generalizar a una superpoblación más amplia de estudios. Una descripción completa de los modelos estadísticos se puede encontrar en Sánchez-Meca et al. (2013).

Tabla 1. Resumen de los Métodos Estadísticos en Meta-Análisis de Generalización de la Fiabilidad

Modelo estadístico	Método de ponderación	Transformación	Autores principales de la propuesta
Efecto fijo	Sin ponderar	Sin transformar	(Vacha-Haase, 1998)
	Inverso de la varianza	Transformación	Hedges y Olkin (1985), Konstantopoulos y Hedges (2009)
Coeficientes variantes	Sin ponderar	Sin transformar	Bonett (2008, 2009, 2010)
Efectos aleatorios	Inverso de la varianza	Transformación	Hedges y Vevea (1998), Raudenbush (2009)
	Inverso de la varianza corregida	Transformación	Hartung (1999), Knapp y Hartung (2003)
	Tamaño muestral	Sin transformar	Hunter y Schmidt (2004)
Transformación adecuada para coeficientes de correlación		Z de Fisher (Fisher, 1921)	
Transformaciones adecuadas para coeficientes alfa		Z de Fisher sobre la raíz cuadrada del coeficiente de fiabilidad (Beretvas, Meyers y Leite, 2002) T de Hakstian-Whalen (Hakstian y Whalen, 1976) L de Bonett (Bonett, 2002)	

El tercer problema se refiere al método de transformación de los coeficientes de fiabilidad. La mayoría de los estudios GF han integrado los coeficientes alfa (Cronbach, 1951) porque son reportados ampliamente en la investigación primaria (López-Pina et al., 2012; Romano & Kromrey, 2009; Scherer & Teo, 2020). Por esta razón, el presente estudio se centra en los coeficientes alfa. Sobre la posibilidad de usar una transformación, algunos autores han recomendado transformar los coeficientes alfa (Feldt & Charter, 2006; Rodriguez & Maeda, 2006; Sawilowsky, 2000), mientras que otros abogan por usar el coeficiente alfa sin transformar (Henson & Thompson, 2002; Leach, et al., 2006; Mason et al., 2007; Thompson & Vacha-Haase, 2000).

En general, al aplicar un método de transformación en estudios GF se siguen dos etapas. En primer lugar, los coeficientes de fiabilidad se transforman a la nueva métrica (por ejemplo, z de Fisher) para realizar los cálculos meta-analíticos. En segundo lugar, tras hacer la estimación promedio de la fiabilidad y el intervalo de confianza, se vuelven a transformar a los coeficientes de fiabilidad a su métrica original para facilitar la interpretación de los resultados. Al transformar los coeficientes de fiabilidad, el método más frecuentemente utilizado ha sido la transformación z de Fisher (α-z) (Sánchez-Meca et al., 2008). Sin embargo, esta transformación, siendo teóricamente apropiada para métodos basados en la administración repetida del test (test-retest) o en la división del test en dos mitades, no lo es para los coeficientes alfa, ya que fue desarrollada para coeficientes de correlación (Fisher, 1921). Para resolver este problema, también se ha aplicado la transformación z de Fisher sobre la raíz cuadrada de los coeficientes alfa (r-z), dado que se puede interpretar como el índice de fiabilidad (Beretvas et al., 2002). Para el coeficiente alfa, una transformación más adecuada que la z de Fisher es la propuesta por Hakstian & Whalen (1976), ya que normaliza la distribución de los coeficientes de fiabilidad. Sin embargo, desde un punto de vista teórico, una transformación aún mejor es la propuesta por Bonett (2002), ya que normaliza la distribución de los coeficientes alfa y estabiliza sus varianzas. Aunque tanto las transformaciones de Hakstian-Whalen (H-W) como las de Bonett (B) son teóricamente más apropiadas para coeficientes alfas que la z de Fisher, no hay demasiados estudios GF que la hayan utilizado (Sánchez-Meca et al., 2008).

Hasta donde el autor conoce solo existe un estudio previo (Sánchez-Meca et al., 2013) que haya analizado el comportamiento de los distintos métodos estadísticos en estudios GF. En ese trabajo los autores compararon las estimaciones promedio de coeficientes alfa junto con el influjo de variables moderadoras usando 13 métodos estadísticos. Por un lado, encontraron ligeras diferencias en cuanto al efecto de la transformación sobre la fiabilidad promedio, que en algunas ocasiones no consiguió normalizar la distribución de coeficientes. Por otro lado, observaron un efecto del tipo de modelo sobre la amplitud de los intervalos de confianza. En concreto, el modelo EF consiguió los intervalos más estrechos.

Asimismo, encontraron diferencias importantes en el análisis de variables moderadoras, de forma que los procedimientos llevaban a conclusiones distintas cuando la relación entre moderador y coeficientes de fiabilidad era fuerte. Los autores concluyen que cuando las

condiciones del modelo EA se cumplen, se emplee la transformación de Bonett sobre los coeficientes alfa junto con la corrección propuesta por Hartung (1999) y Knapp y Hartung (2003).

Objetivos del Estudio

Este estudio pretende continuar el trabajo iniciado por Sánchez-Meca et al. (2013) usando dos procedimientos estadísticos adicionales en tres conjuntos de datos reales. Para ello, se llevó a cabo un estudio GF del Maslach Burnout Inventory (MBI; Maslach & Jackson, 1981), utilizando el coeficiente alfa como estimador de la fiabilidad. Se usaron los datos recogidos por Aguayo et al. (2011), con un total de 51 muestras independientes provenientes de 45 estudios primarios. Dado que el MBI consta de tres subescalas diferentes (cansancio emocional, despersonalización y realización personal), se realizaron cálculos meta-analíticos independientes para cada una de ellas. El propósito general de este trabajo es comparar los resultados obtenidos con diferentes métodos estadísticos para realizar un meta-análisis GF. Los objetivos específicos fueron: (a) comparar la forma de las distribuciones de los coeficientes alfa originales y transformados; (b) examinar las estimaciones promedio de fiabilidad y los intervalos de confianza obtenidos con 15 métodos estadísticos; y (c) analizar el comportamiento de los métodos seleccionados para evaluar la influencia de las características muestrales que puedan explicar la variabilidad en los coeficientes de fiabilidad.

Método

Se realizaron 15 comparaciones, para cada una de las tres dimensiones del MBI, con los procedimientos estadísticos seleccionados: (a) el método de mínimos cuadrados ordinarios (MCO) aplicado sobre los coeficientes alfa no transformados; (b) el modelo de efecto fijo (EF) aplicado sobre la transformación z de Fisher de los coeficientes alfa (α-z) y de los coeficientes índice de fiabilidad (r-z), así como sobre las transformaciones de Hakstian-Whalen (H-W) y de Bonett (B); (c) el modelo de coeficientes variantes (CV) sobre los coeficientes alfa sin transformar; (d) los modelos de efectos aleatorios (EA) y de efectos aleatorios corregidos (EA-c) aplicados sobre las transformaciones mencionadas anteriormente; y (e) la aproximación de Hunter y Schmidt, que usa el modelo de efectos aleatorios ponderando los coeficientes alfa sin transformar por el tamaño muestral (RE-n).

Nótese que todos los procedimientos estadísticos seleccionados son coherentes con las propuestas teóricas de sus correspondientes autores. Por ejemplo, en la propuesta original de Vacha-Haaase (1998) el método MCO se aplicó sobre coeficientes alfa sin transformar. Los modelos de EF, EA y EA-c se propusieron para ser usados junto con alguna transformación de los coeficientes alfa (Hedges & Olkin, 1985). Hunter y Schmidt (2004) introdujeron el modelo RE-n para el que no recomendaban usar transformaciones. Finalmente, de acuerdo con Bonett (2010), el modelo CV no se debe usar combinado con alguna transformación.

El MBI (Maslach & Jackson, 1981) cuenta con 22 ítems que son valorados en frecuencia, con un formato de respuesta tipo Likert de siete alternativas de respuesta. Las subescalas cansancio emocional, despersonalización y realización personal tienen, respectivamente nueve, cinco y ocho ítems.

Para cada subescala del MBI y procedimiento seleccionado, se hicieron estimaciones de la fiabilidad promedio y se calculó un intervalo de confianza. Asimismo, se calculó un índice de discrepancia para comparar las estimaciones de la fiabilidad promedio, tomando como referencia el promedio no ponderado de los coeficientes no transformados (método de MCO). Se compararon los intervalos de confianza en términos de sus amplitudes. La influencia de las características del estudio (variables moderadoras) se comparó con valores *p* y con índices de proporción de varianza explicada. La descripción de estos estadísticos se puede encontrar en Sánchez-Meca et al. (2013).

Finalmente, con el fin de comparar la forma de las distribuciones de los coeficientes alfa transformados y no transformados, se calcularon estadísticos descriptivos de tendencia central, variabilidad y forma. Se consideró que el grado de asimetría y curtosis era elevado a partir de valores mayores al rango [-2,2] (Lloret-Segura et al., 2014). Para valorar si las distribuciones de coeficientes se ajustaban a la normalidad se usó la

Tabla 2. Estadísticos Descriptivos para los Coeficientes Alfa Transformados y no Transformados (k=51)

	Cansancio Emocional					Despersonalización					Realización Personal				
	Alfa	α-*z*	*r*-*z*	H-W	B	Alfa	α-*z*	*r*-*z*	H-W	B	Alfa	α-*z*	*r*-*z*	H-W	B
Min.	.660	.793	1.134	.368	-2.996	.430	.460	.785	.554	-1.772	.490	.536	.867	.391	-2.813
Max.	.950	1.832	2.178	.698	-1.079	.830	1.188	1.534	.829	- .562	.940	1.738	2.085	.799	- .673
Rango	.290	1.039	1.044	.330	1.917	.400	.728	.749	.275	1.210	.450	1.202	1.218	.408	2.140
Media	.870	1.362	1.707	.500	-2.098	.697	.883	1.225	.665	-1.239	.763	1.029	1.373	.612	-1.492
Mediana	.880	1.376	1.722	.493	-2.120	.720	.908	1.251	.654	-1.273	.780	1.045	1.390	.604	-1.514
DT	.049	.183	.183	.058	.340	.093	.178	.182	.067	.301	.080	.189	.192	.067	.334
CV	.056	.134	.107	.116	- .162	.133	.202	.148	.101	- .243	.105	.184	.140	.109	- .224
Asimetría	-1.784	- .377	- .388	.753	.265	- .669	- .216	- .243	.291	.117	-1.380	.342	.290	.169	- .644
Curtosis	4.917	1.010	1.032	1.529	.860	- .130	- .799	- .759	- .744	- .897	3.102	3.193	3.148	2.313	3.726
Test S-W	.000	.282	.268	.051	.403	.027	.360	.350	.267	.391	.000	.001	.001	.004	.001

Nota. Min and Max=Valores mínimos y máximos de las distribuciones; *CV*=Coeficiente de variación; test S-W=Valor *p* obtenido con el test de Shapiro-Wilk; Alfa=Coeficiente alfa no transformado; α-*z*=Transformación *z* de Fisher del coeficiente alfa; *r*-*z*=Transformación *z* de Fisher de la raíz cuadrada del coeficiente alfa; H-W=Transformación de Hakstian-Whalen; B=Transformación de Bonett.

prueba de Shapiro-Wilk dado que el tamaño muestral era bajo. Los análisis estadísticos se realizaron con el software R (versión 4.2.2), utilizando los paquetes metafor (versión 3.8-1; Viechtbauer, 2010), psych (versión 2.2.9; Revelle, 2023) y nortest (versión 1.0-4; Gross, 2012).

Resultados

El primer objetivo fue comparar la forma de las distribuciones de los coeficientes alfa y las cuatro transformaciones utilizadas en el presente trabajo. La Tabla 2 presenta los estadísticos descriptivos principales para las tres subescalas del MBI: cansancio emocional (CE), despersonalización (D) y realización personal (RP). Los coeficientes alfa originales mostraron valores de asimetría y curtosis pronunciados, lo que llevó al rechazo de la hipótesis nula de normalidad en las tres subescalas. En CE y D, las cuatro transformaciones consiguieron alcanzar la normalidad de las distribuciones, mientras que ninguna de ellas lo consiguió en RP. Las transformaciones H-W y B lograron la mayor reducción de asimetría y curtosis, aunque no en todas las situaciones. En CE, la transformación B mostró un mejor ajuste a la normalidad, aunque la mayoría de las transformaciones redujeron la asimetría y la curtosis a valores bajos. La transformación de B también mostró el mejor ajuste a la normalidad en D. Sin embargo, todas las transformaciones aumentaron ligeramente la curtosis de las distribuciones en esta subescala. En RP, la transformación H-W tuvo un rendimiento mejor que las otras transformaciones, aunque no cumplió con la normalidad. Ninguna de las transformaciones logró reducir con éxito la curtosis en esta subescala.

El segundo objetivo fue comparar los coeficientes de fiabilidad promedio e intervalos de confianza obtenidos con cada uno de los 15 métodos estadísticos. La Tabla 3 muestra los resultados para las tres subescalas del MBI. Tomando la media no ponderada (modelo de mínimos cuadrados ordinarios) de los coeficientes alfa no transformados como referencia, la mayoría de las estimaciones de la fiabilidad promedio en la dimensión CE mostraron discrepancias menores al 1.5%. El método de EF obtuvo discrepancias mayores que los métodos MCO y EA en las estimaciones promedio. Específicamente, la mayor discrepancia se alcanzó con el modelo de EF combinado con la transformación H-W (1.41%), lo que llevó a un cambio en el valor alfa de .870 a .882.

Las mayores discrepancias se obtuvieron en la dimensión D, siendo algunas de ellas negativas. En este caso, el modelo de EA-n logró la mayor discrepancia (-6.12), lo que llevó a una reducción en el valor alfa de .671 a .654. Los métodos de EF nuevamente produjeron discrepancias elevadas en la dimensión RP, siendo la mayor la del modelo de EF combinado con la transformación H-W (2.88%), que llevó a un aumento en los valores alfa de .763 a .785. En cuanto a la amplitud de los intervalos de confianza, hubo una clara influencia de los modelos estadísticos en lugar del procedimiento de transformación. En general, los métodos de EF produjeron los intervalos más estrechos, seguidos del modelo de CV, el de EA-n, los de RE y RE-c, y, finalmente, el de MCO, que mostró los intervalos más anchos en dos de las tres situaciones.

El tercer objetivo fue comparar los métodos seleccionados al evaluar la influencia de las

Tabla 3. Resultados de las Estimaciones Alfa Promedio y sus Intervalos de Confianza

Modelo estadístico	Método de transformación	Cansancio Emocional					Despersonalización					Realización Personal				
		$\bar{\alpha}$	D	Li	Ls	A	$\bar{\alpha}$	D	Li	Ls	A	$\bar{\alpha}$	D	Li	Ls	A
MCO	Alfa	.870	--	.856	.884	.028	.697	--	.671	.723	.052	.763	--	.741	.785	.045
EF	α-z	.878	.93	.875	.881	.006	.666	-4.52	.659	.672	.014	.772	1.17	.767	.777	.010
EF	r-z	.878	.93	.875	.881	.006	.665	-4.63	.658	.672	.014	.772	1.15	.767	.777	.010
EF	H-W	.882	1.41	.880	.884	.004	.683	-1.95	.677	.690	.012	.785	2.88	.781	.789	.008
EF	B	.878	.96	.876	.880	.005	.668	-4.10	.662	.675	.013	.773	1.32	.769	.777	.008
CV	Alfa	.870	--	.866	.874	.008	.697	--	.689	.707	.018	.763	--	.756	.772	.015
EA	α-z	.876	.77	.867	.886	.019	.706	1.34	.679	.732	.053	.776	1.66	.755	.794	.039
EA	r-z	.876	.76	.867	.886	.019	.706	1.27	.678	.731	.053	.775	1.62	.755	.794	.039
EA	H-W	.876	.72	.867	.885	.018	.706	1.28	.678	.732	.053	.774	1.42	.751	.795	.044
EA	B	.877	.83	.867	.886	.019	.709	1.71	.682	.733	.051	.777	1.79	.756	.795	.039
EA-c	α-z	.876	.77	.864	.888	.023	.706	1.34	.680	.731	.051	.776	1.66	.754	.796	.042
EA-c	r-z	.876	.76	.864	.888	.024	.706	1.27	.679	.731	.051	.775	1.62	.753	.796	.042
EA-c	H-W	.876	.72	.864	.888	.018	.706	1.28	.680	.731	.051	.774	1.42	.753	.794	.041
EA-c	B	.877	.83	.865	.888	.019	.709	1.71	.683	.733	.050	.777	1.79	.755	.797	.042
EA-n	Alfa	.874	.49	.865	.884	.019	.654	-6.12	.628	.681	.053	.764	.16	.747	.781	.033

Nota. $\bar{\alpha}$ =Coeficiente alfa promedio (re-transformado cuando ha sido necesario); D=Índice de discrepancia (%); Li y Ls=Límites inferior y superior del intervalo de confianza (1 – α=.95); A=Amplitud del intervalo de confianza; MCO=Mínimos cuadrados ordinarios; EF=Modelo de efecto fijo; EA=Modelo de efectos aleatorios; EA-c=Modelo de efectos aleatorios corregido; RE-n=Modelo de efectos aleatorios de Hunter y Schmidt (2004); Alfa=Coeficiente alfa no transformado; α-z=Transformación z de Fisher del coeficiente alfa; r-z=Transformación z de Fisher de la raíz cuadrada del coeficiente alfa; H-W=Transformación de Hakstian-Whalen; B=Transformación de Bonett.

Tabla 4. Resultados del Análisis de Moderadores

Modelo estadístico	Método de transformación	Cansancio Emocional				Despersonalización				Realización Personal			
		Versión		DT		Versión		DT		Versión		DT	
		p	R2	p	R2	p	R2	p	R2	p	R2	p	R2
MCO	Alfa	.098	.036	.000	.524	.081	.043	.059	.090	.869	0	.039	.112
EF	α-z	.000	0	.000	.009	.000	.078	.000	.338	.001	0	.002	.434
EF	r-z	.000	0	.000	.009	.000	.078	.000	.337	.000	0	.002	.422
EF	H-W	.000	0	.000	0	.000	.063	.000	.388	.215	0	.001	.620
EF	B	.000	0	.000	0	.000	.075	.000	.335	.000	0	.000	.475
CV	Alfa	.000	.050	.000	.426	.000	.050	.000	.026	000	.050	.173	0
EA	α-z	.048	0	.000	.009	.098	.078	.107	.110	.924	0	.695	0
EA	r-z	.048	0	.000	.009	.097	.078	.106	.144	.931	0	.695	0
EA	H-W	.033	0	.000	0	.104ár	.063	.024	.388	.979	0	.011	.620
EA	B	.037	0	.000	0	.099	.075	.025	.335	.891	0	.015	.475
EA-c	α-z	.099	0	.000	.009	.093	.078	.033	.338	.927	0	.043	.434
EA-c	r-z	.099	0	.000	.009	.092	.078	.034	.337	.934	0	.043	.422
EA-c	H-W	.086	0	.000	0	.090	.063	.033	.388	.977	0	.046	.620
EA-c	B	.086	0	.000	0	.097	.075	.033	.335	.894	0	.039	.475
EA-n	Alpha	.098	.036	.000	.524	.081	.043	.059	.090	.869	0	.039	.112

Nota. Versión=Versión del test (original y adaptado); DT=Desviación típica de las puntuaciones de la subescala; *p*=Valor *p* del contraste de hipótesis (cuando la prueba t fue usada, se obtuvo a partir de una distribución t-Student con k-2 grados de libertad); R2=Proporción de varianza explicada por el moderador en el análisis de regresión simple (se truncó a cero cuando fue negativo); MCO=Mínimos cuadrados ordinarios; EF=Modelo de efecto fijo; EA=Modelo de efectos aleatorios; EA-c=Modelo de efectos aleatorios corregido; RE-n=Modelo de efectos aleatorios de Hunter y Schmidt (2004); Alfa=Coeficiente alfa no transformado; α-z=Transformación z de Fisher del coeficiente alfa; r-z=Transformación z de Fisher de la raíz cuadrada del coeficiente alfa; H-W=Transformación de Hakstian-Whalen; B=Transformación de Bonett.

características de los estudios primarios que pueden explicar la variabilidad en los coeficientes alfa. Para este propósito se eligió un predictor dicotómico y otro continuo. El primero fue la versión del test (original y adaptada a otros idiomas), y el segundo fue la desviación típica de las puntuaciones en cada subescala del MBI. La Tabla 4 muestra los resultados de los análisis de moderadores, informando los valores y los índices R2 para cada modelo de regresión lineal.

Los contrastes de hipótesis no mostraron acuerdo entre los métodos para ninguno de los dos predictores. En general, los valores *p* más pequeños para el predictor dicotómico fueron obtenidos por los métodos de EF y de CV, lo que llevaba al rechazo de la hipótesis nula de no efecto, seguidos por los métodos de MCO, EA y EA-n, que no rechazaron la hipótesis nula. Los métodos de EA-c proporcionaron resultados más conservadores que los métodos de EA en una de las tres situaciones, aunque ambos métodos llegaron a las mismas conclusiones en la mayoría de las ocasiones. Un resultado interesante fue que, en dos de las tres situaciones, EF y CV llevaron al rechazo de la hipótesis nula, mientras que el resto de los métodos concluyeron lo contrario.

Los métodos de EF y CV también proporcionaron los valores *p* más pequeños para el predictor continuo, seguidos por la combinación del modelo de EA y las transformaciones H-W y B, y por los métodos de MCO, EA-n y AE-c, que arrojaron resultados similares. La combinación del modelo de EA y las transformaciones z de Fisher (tanto α-z como r-z) proporcionó los resultados más conservadores, sin llegar a rechazar la hipótesis nula en dos de las tres situaciones. En este caso, es interesante señalar que, en dos de las tres situaciones, la combinación de EA y las transformaciones z de Fisher llevó al rechazo de la hipótesis nula, mientras que se llegó a la conclusión opuesta con el resto de los métodos.

Los índices de proporción de varianza explicada mostraron concordancia para el predictor dicotómico, pero no para el continuo. Para el primero, todos los métodos arrojaron porcentajes muy cercanos a cero. En cambio, para el segundo, hubo una influencia del procedimiento de transformación, ya que, en la mayoría de las situaciones, los coeficientes no transformados obtuvieron valores muy diferentes a los transformados. En esta línea, las transformaciones B y H-W arrojaron los mismos resultados independientemente del modelo estadístico. Un resultado inesperado fue que la combinación del modelo de EA y las transformaciones z de Fisher difería considerablemente de los demás métodos de ponderación en dos de las tres situaciones. No hubo discrepancias entre las dos transformaciones z de Fisher ni para los predictores dicotómicos ni para los continuos.

Discusión

Los estudios meta-analíticos de Generalización de la Fiabilidad (GF) tratan de proporcionar una estimación de la fiabilidad promedio de un test y, en caso de encontrar heterogeneidad elevada, explicarla a partir de variables moderadoras. Los métodos estadísticos que se usan en este tipo de estudios difieren en tres aspectos: el modelo estadístico, la forma de ponderación y el tipo de transformación de los coeficientes de fiabilidad. Actualmente no hay unos criterios consensuados para elegir entre los distintos métodos existentes y los investigadores pueden optar arbitrariamente por cualquiera de ellos (Sánchez-Meca et al., 2013). Ante esta situación es lógico preguntar si el uso de distintos métodos conlleva resultados o conclusiones diferentes. Es decir, si las conclusiones a las que se llega tras un meta-análisis GF pueden cambiar en función del método estadístico escogido. Por ello, el objetivo general de este trabajo consistió en analizar 15 métodos estadísticos que se usan en la realización de estudios GF empleando datos reales obtenidos tras la aplicación del Maslach Burnout Inventory (MBI; Maslach & Jackson, 1981) provenientes de 45 estudios primarios y 51 muestras independientes recogidas por Aguayo et al. (2011). Este estudio pretende sumarse al trabajo iniciado por Sánchez-Meca et al. (2013), añadiendo dos procedimientos adicionales para las comparaciones.

En concreto, el primer objetivo fue comparar la distribución de coeficientes alfa y transformados. Los resultados mostraron que los coeficientes alfa obtuvieron valores de asimetría y curtosis moderados, que llevaron al rechazo de la hipótesis de normalidad de las distribuciones. Todas las transformaciones utilizadas consiguieron mejorar los valores de asimetría de forma considerable, llevando en la mayoría de las ocasiones a normalizar las distribuciones. Un resultado interesante fue que en la mayoría de las veces los valores de curtosis aumentaron tras usar las transformaciones. En este sentido, la transformación de Hakstian-Whalen fue la que produjo incrementos más pequeños. Estos resultados coinciden con los encontrados por Sánchez-Meca et al. (2013).

Como cabía esperar según la teoría estadística, algunas transformaciones consiguieron resultados más similares entre sí que otras. Por una parte, las dos transformaciones basadas en la z de Fisher (una aplicada sobre los coeficientes alfa y otra sobre su raíz cuadrada) consiguieron valores promedios de fiabilidad y amplitudes en los intervalos de confianza prácticamente iguales entre sí, y distintos al resto de transformaciones. Este resultado coincide con lo encontrado en el estudio de simulación de Feldt y Charter (2006). Por otra parte, algo parecido ocurrió con las dos transformaciones que se han propuesto específicamente para meta-análisis GF (la de Hakstian-Whalen & Bonett).

El segundo objetivo consistió en estudiar las discrepancias entre los valores de alfa promedio a

partir de coeficientes transformados y sin transformar. Las mayores discrepancias estuvieron entre el 1.41% y el 6.12% para las distintas situaciones. En general, estas discrepancias no implicaron conclusiones distintas en cuanto a la fiabilidad media de las escalas, ya que en la mayoría de las dimensiones del MBI estaba por encima del valor habitualmente recomendado .70 (Nunnally & Berstein, 1978). Incluso los límites inferiores de los intervalos de confianza mostraron valores por encima de este punto de corte. No obstante, la dimensión despersonalización obtuvo valores alfa promedio que oscilaban por debajo y por encima de .70. Así, dependiendo del método que se use, un investigador aplicado podría haber concluido que la escala está o no está por encima de los valores de fiabilidad habitualmente recomendados. En concreto, los métodos que consiguieron un alfa promedio mayor a .70 fueron el modelo de efectos aleatorios y el modelo de efectos aleatorios corregido (Knapp & Hartung, 2003) combinado con alguna transformación de coeficientes alfa.

Estos resultados coinciden con las tendencias encontradas por Sánchez-Meca et al. (2013), donde la mayoría de las discrepancias fueron menores del 5%. Además, en ambos estudios las mayores discrepancias tendían a obtenerse con el modelo de efecto fijo, especialmente, cuando se combinaba con la transformación de Hakstian-Whalen. En cuanto a los intervalos de confianza, los resultados son los esperados de acuerdo con los modelos de meta-análisis. La amplitud de los intervalos vino determinada por el tipo de ponderación de coeficientes y no por el método de transformación. En concreto, la amplitud de los intervalos fue menor con el modelo de efecto fijo, ya que solo incluye en su ponderación la varianza intra-estudios (Borenstein et al., 2015). El modelo de Coeficientes Variantes de Bonett (2002) consiguió unos resultados muy similares al anterior, ya que se puede considerar un tipo de modelo de efecto fijo (Sánchez-Meca et al., 2013).

Entorno a los valores promedio de fiabilidad se encontró altos niveles de heterogeneidad, superiores al 90% (Aguayo et al., 2011). Por ello, el último objetivo del presente estudio consistió en analizar la influencia de variables moderadoras que pudieran explicar tal heterogeneidad. De acuerdo con la teoría estadística (Sánchez-Meca et al., 2013) se pueden realizar ciertas hipótesis. En primer lugar, dado que las fuentes de varianza del modelo de efecto fijo son menores que las del modelo de efectos aleatorios, será más probable rechazar la hipótesis nula de no efecto del moderador con el primer modelo más que con el segundo.

Los resultados del presente estudio confirman lo hipotetizado por la teoría, ya que con el modelo de efecto fijo los p-valores fueron más bajos que con los del modelo de efectos aleatorios en casi todos los contrastes de hipótesis, lo que conllevaba un rechazo de la hipótesis nula en mayor cantidad de ocasiones.

En segundo lugar, las discrepancias entre los métodos estadísticos deberían aparecer solamente cuando la influencia del moderador es moderada. Con una variable moderadora cuya influencia fuera muy fuerte o débil los métodos estadísticos deberían coincidir en sus resultados. Los resultados de este estudio no están en línea con estas hipótesis. De los dos moderadores utilizados, cabe esperar que la desviación típica de las puntuaciones tenga una influencia fuerte a la hora de explicar la variabilidad en los coeficientes alfa promedio. Esto se deriva de la Teoría Clásica de los Tests que indica que la fiabilidad de un test depende de la variabilidad de las puntuaciones.

Teniendo en cuenta lo anterior, no cabía esperar, por un lado, diferencias entre los distintos métodos estadísticos con la variable moderadora desviación típica de las puntuaciones del test. Sin embargo, mientras que con el modelo de efectos aleatorios se mantenía la hipótesis nula de no efecto en dos de tres situaciones, con el modelo de efectos aleatorios corregido se rechazó en todas las ocasiones.

En principio, los resultados obtenidos con esta corrección, propuesta por Hartung (1999) y Knapp y Hartung (2003), son menos sesgados que sin ella, ya que está basada en el hecho de que las varianzas muestrales se calculan en el modelo original como si fuesen conocidas cuando en la práctica tienen que ser estimadas.

Por otro lado, como se esperaba, cuando el efecto del moderador era bajo, como sucedió con la variable moderadora versión del test, hubo diferencias entre métodos estadísticos. El modelo de efectos aleatorios obtuvo resultados en general menos conservadores que los modelos de efectos

aleatorios corregido y el de Hunter y Schmidt (2004), que pondera los coeficientes por el tamaño muestral, pudiendo llevar a rechazar equivocadamente la hipótesis nula.

En cuanto a las transformaciones, dado que no intervienen en el cómputo del contraste de hipótesis, no se esperaban diferencias dentro del mismo modelo de meta-análisis. Los resultados apoyaron esta hipótesis ya que no hubo demasiada influencia de éstas sobre el rechazo de las hipótesis nulas ni sobre el cómputo del tamaño del afecto del modelo. Estos resultados coinciden con lo obtenido por Sánchez-Meca et al. (2013).

Limitaciones y futuros estudios

El presente estudio cuenta con varias limitaciones. En primer lugar, los resultados se ven necesariamente restringidos a las situaciones analizadas, no siendo generalizables ni a otros estudios primarios ni a otros instrumentos de medición psicológica.

En segundo lugar, el análisis de los métodos estadísticos en muestras reales puede ser muy útil para estudiar su comportamiento en situaciones aplicadas donde se producen casuísticas que a menudo se escapan de los limitados diseños de los estudios de simulación. Como dijeron Feldt y Charter (2006), los investigadores aplicados no trabajan en un mundo ideal. Por ello, sería deseable que se realizaran más estudios de comparación de métodos con otros tests psicológicos. Sin embargo, aún más importante es realizar estudios de simulación de los que se puedan extraer recomendaciones acerca del rendimiento de los diferentes estadísticos existentes en la literatura.

Finalmente, este trabajo solamente recoge coeficientes alfa. A pesar de ser menos frecuentes en la literatura psicométrica, existen muchos otros coeficientes de fiabilidad como omega, test-retest o KR-21. Por ello, futuros estudios podrían analizar la influencia de los distintos métodos estadísticos de meta-análisis en estos coeficientes de fiabilidad.

Recomendaciones y Conclusiones

En este estudio sobre métodos estadísticos en meta-análisis de generalización de fiabilidad se han comparado un total de 15 procedimientos estadísticos, obtenidos a partir de la combinación de 5 tipos de transformaciones (incluido el coeficiente alfa sin transformar) y 6 modelos meta-analíticos. En base a estos resultados y los obtenidos en investigaciones previas (Sánchez-Meca et al., 2012) y estudios de simulación (Bonett, 2010; López-López et al., 2013; López-Pina et al., 2012), se pueden derivar algunas conclusiones y recomendaciones generales.

Las transformaciones tienden a conseguir la normalización de los coeficientes de fiabilidad, aunque es recomendable comprobarlo en cada aplicación ya que no siempre lo consiguen. Aquellas transformaciones sugeridas en el ámbito del meta-análisis de generalización de fiabilidad, en concreto las propuestas por Hakstian y Whalen (1976) y Bonett (2002; 2010), parecen comportarse de manera más similar entre sí y de forma distinta a otras (z de Fisher). Por ello, se recomienda usar alguna de las primeras, y particularmente la de Bonett, ya que tiene en cuenta que el supuesto de ítems paralelos del coeficiente alfa no suele cumplirse (Bonett, 2002; López-López et al., 2013; López-Pina et al., 2012; Ondé & Alvarado, 2022; Paniagua et al., 2022).

Los métodos estadísticos obtienen resultados similares en cuanto a la estimación de la fiabilidad promedio (discrepancias menores al 7%). No obstante, estas pequeñas diferencias pueden repercutir en la interpretación del coeficiente de fiabilidad cuando se sitúan en torno a los puntos de corte que se usan para valorar cualitativamente la fiabilidad de la escala en una aplicación concreta (e.g., con valores de fiabilidad sobre .70, .80 o .90). Uno de los métodos que consiguió fiabilidades promedio más altas fue el de efectos aleatorios corregido (Hartung, 1999; Knapp & Hartung, 2003), que es el que se suele recomendar habitualmente (Borenstein et al., 2015; Sánchez-Meca et al., 2013).

Hay que tener en cuenta que, con el modelo de efectos aleatorios (tanto el no corregido como el corregido), el intervalo de confianza suele ser más amplio que con el modelo de efecto fijo. Por ello, es más fácil conseguir un efecto estadísticamente significativo de las variables moderadoras con este último que con el primero. Sin embargo, esto no debe ser un criterio de elección de uno sobre el otro. El modelo de efecto fijo solamente se debe utilizar cuando se quiera limitar la generalización de la fiabilidad promedio al conjunto de estudios analizados, o muy similares (Borenstein et al., 2015; Sánchez-Meca et al., 2013).

Finalmente, se produjeron diferencias entre los métodos estadísticos en el análisis de variables moderadoras tanto cuando el efecto de la variable era débil como cuando era fuerte. Ante esta discrepancia, y a falta de estudios de simulación sobre este problema, se recomienda utilizar el método de efectos aleatorios corregido, ya que ha obtenido los mejores resultados en estudios de simulación (López-López et al., 2013) y es más apropiado en la mayoría de las ocasiones (Borenstein et al., 2015; Field, 2005; Sánchez-Meca & López-Pina, 2008; Sánchez-Meca et al., 2013; Schmidt, 2010).

Referencias

Aguayo, R., Vargas, C., de la Fuente, E. I., & 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.

Beretvas, S. N., Meyers, J. L., & Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement, 62*, 570-589. https://doi.org/10.1177/0013164402062004003

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*, 335-340.

Bonett, D. G. (2008). Meta-analytic interval estimation for bivariate correlations. *Psychological Methods, 13*, 173-181. https://doi.org/10.1037/a0012868

Bonett, D. G. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods, 14*, 225-238. https://doi.org/10.1037/a0016619

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods, 15*, 368-385. https://doi.org/10.1037/a0020142

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2015*). Introduction to meta-analysi*s (2nd ed.). Wiley.

Botella, J., & 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., Suero, M., & Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods, 15*, 386-397. https://doi.org/10. 1037/a0019626

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

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

Feldt, L. S., & Charter, R. A. (2006). Averaging internal consistency reliability coefficients. *Educational and Psychological Measurement, 66*, 215-227. https://doi.org/10.1177/0013164404273947

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods, 10*, 444–467. https://doi.org/10.1037/1082-989X.10.4.444

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron, 1*, 3–32.

Gronlund, N. E., & Linn, R. L. (1990). Measurement and evaluation in teaching. (6th ed.). Macmillan.

Gross, J. (2023). Nortest: Test for normality. R package version 1.0-4. Disponible en https://CRAN.R-project.org/package=nortest

Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika, 41*, 219–231. https://doi:10.1007/BF02291840

Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal, 41*, 901-916. https://doi.org/10.1002/(SICI)1521-4036(199912)41:83.0.CO;2-W

Hedges, L. V., & Olkin, I. (1985). Statistical methods in meta-analysis. Academic Press.

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

Henson, R. K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting “reliability generalization” studies.

Measurement and Evaluation in Counseling and Development, 35, 113-126.

Hunter, J. E., & Schmidt, F. L. (2004). Methods of meta-analysis: Correcting errors and bias in research findings (2nd ed.). Sage.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine, 22*, 2693-2710. https://doi:10.1002/sim.1482

Konstantopoulos, S., & Hedges, L. V. (2009). Analyzing effect sizes: Fixed-effects models. En H. Cooper, L. V. Hedges & J. C. Valentine, (Eds.). *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 279-293). Russell Sage Foundation.

Laird, N. M., & Mosteller, F. (1990). Some statistical methods for combining experimental results. International *Journal of Technology Assessment in Health Care, 6*, 5-30.

https://doi.org/10.1017/S0266462300008916

Leach, L. F., Henson, R. K., Odom, L. R., & Cagle, L. S. (2006). A reliability generalization study of the Self-Description Questionnaire. *Educational and Psychological Measurement, 66*, 285–304. https://doi.org/10.1177/0013164405284030

Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., & Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: Una guía práctica, revisada y actualizada. *Anales de Psicología, 30*(3), 1151–1169. https://doi.org/10.6018/analesps.30.3.199361

López-López, J. A., Botella, J., Sánchez-Meca, J., & Marín-Martínez, F. (2013). Alternatives for Mixed-Effects Meta-Regression Models in the Reliability Generalization Approach: A simulation study. *Journal of Educational and Behavioral Statistics, 38*, 443–469. https://doi.org/10.3102/1076998612466142

López-Pina, J. A., Sánchez-Meca, J., & López-López, J. A. (2012). Methods for averaging alpha coefficients in reliability generalization studies. *Psicothema, 24*, 161-166.

Maslach, C., & Jackson, S. E. (1981). The measurement of experienced burnout. *Journal of Organizational Behavior, 2*, 99-113.

Mason, C., Allam, R., & Brannick, M. T. (2007). How to meta-analyze coefficient-of-stability estimates: Some recommendations on Monte Carlo studies. *Educational and Psychological Measurement, 67*, 765-783. https://doi.org/10.1177/0013164407301532

Nunnally, J. C., & Bernstein, I. H. (1978). Psychometric testing. McGraw-Hill.

Ondé, D., & Alvarado, J. M. (2022). Contribución de los modelos factoriales confirmatorios a la evaluación de estructura interna desde la perspectiva de la validez. *Revista Iberoamericana de Diagnóstico y Evaluación – e Avaliação Psicológica, 66,* 5-21. https://doi.org/10.21865/RIDEP66.5.01

Paniagua, D., Alvarado, J. M., Olivares, M., Ruiz, I., Romero-Suárez, M., & Aguayo-Estremera, R. (2022). Estudio de Seguimiento de las Recomendaciones sobre Análisis Factorial Exploratorio en RIDEP. *Revista Iberoamericana de Diagnóstico y Evaluación – e Avaliação Psicológica, 66*, 127-139. https://doi.org/10.21865/RIDEP66.5.10

Paniagua, D., Sánchez-Iglesias, I., Miguel-Álvaro, A., Casas-Aragonez, N., Aparicio-García, M. A., & Aguayo-Estremera, R. (2022). Prácticas cuestionables en estudios de validez de instrumentos de medición psicológica: Comunalidades y unicidades de la crisis de replicabilidad en el campo de la psicometría. *Revista Iberoamericana de Diagnóstico y Evaluación – e Avaliação Psicológica, 66*, 23-34. https://doi.org/10.21865/RIDEP66.5.02

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. En H. Cooper, L. V. Hedges & J. C. Valentine, (Eds.*). The handbook of research synthesis and meta-analysis* (2nd ed., pp. 295-315). Russell Sage Foundation.

Revelle, W. (2023). psych: Procedures for psychological, psychometric, and personality research. Northwestern University, Evanston, Illinois, USA. R package version 2.3.9. Disponible en https://CRAN.R-project.org/package=psych

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*, 306-322. https://doi.org/10.1037/1082-989X.11.3.306

Romano, J. L., & Kromrey, J. D. (2009). What are the consequences if the assumption of independent observations is violated in reliability generalization meta-analysis studies? *Educational and Psychological*

Measurement, 69, 404-428. https://doi.org/10.1177/0013164408323237

Rouse, S. V. (2007). Using reliability generalization methods to explore measurement error: An illustration using the MMPI-2 PSY-5 scales. *Journal of Personality Assessment, 88,* 264-275. https://doi.org/10.108/00223890701293908

Sánchez-Meca, J., & 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., & López-López, J. A. (2008). Una revisión de los estudios meta-analíticos de generalización de la fiabilidad. *Escritos de Psicología,* 2-1, 110-121.

Sánchez-Meca, J., López-López, J. A., & López-Pina, J. A. (2013). Some recommended statistical analytic practices when reliability generalization (RG) studies are conducted. *British Journal of Mathematical and Statistical Psychology, 66,* 402-425. https://doi.org/10.1111/j.2044-8317.2012.02057.x

Sawilowsky, S. S. (2000). Psychometrics versus datametrics: Comment on Vacha-Haase's 'reliability generalization' method and some EPM editorial policies. *Educational and Psychological Measurement, 60,* 157–173. https://doi.org/10.1177/00131640021970439

Scherer, R., & Teo, T. (2020). A tutorial on the meta-analytic structural equation modeling of reliability coefficients. *Psychological Methods, 25,* 747–775. https://doi.org/10.1037/met0000261

Schmidt, F. L. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science, 5,* 233–242. https://doi.org/10.1177/174569161036933

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60,* 174–195. https://doi.org/10.1177/0013164400602002

Vacha-Haase T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58,* 6-20. https://doi.org/10.1177/00131644980580010 02

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development, 44,* 159-168. https://doi.org/10.1177/0748175611409845

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1-48. https://doi.org/10.18637/jss.v036.i03