anales de psicología, 2017, vol. 33, nº 3 (octubre), 755-782
http://dx.doi.org/10.6018/analesps.33.3.268401

ISSN print edition: 0212-9728. ISSN web edition (http://revistas.um.es/analesps): 1695-2294

Un viaje alrededor de alfa y omega para estimar la fiabilidad de consistencia interna

Carme Viladrich*, Ariadna Angulo-Brunet y Eduardo Doval

Universitat Autònoma de Barcelona (España).

Resumen: En este trabajo se presenta una guía conceptual y práctica para estimar la fiabilidad de consistencia interna de medidas obtenidas mediante suma o promedio de ítems con base en las aportaciones más recientes de la psicometría. El coeficiente de fiabilidad de consistencia interna se presenta como un subproducto del modelo de medida subyacente en las respuestas a los ítems y se propone su estimación mediante un procedimiento de análisis de los ítems en tres fases, a saber, análisis descriptivo, comprobación de los modelos de medida pertinentes y cálculo del coeficiente de consistencia interna y su intervalo de confianza. Se proporcionan las siguientes fórmulas: (a) los coeficientes alfa de Cronbach y omega para medidas unidimensionales con ítems cuantitativos (b) los coeficientes omega ordinal, alfa ordinal y de fiabilidad no lineal para ítems dicotómicos y ordinales, y (c) los coeficientes omega y omega jerárquico para medidas esencialmente unidimensionales con efectos de método. El procedimiento se generaliza al análisis de medidas obtenidas por suma ponderada, de escalas multidimensionales, de diseños complejos con datos multinivel y/o faltantes y también al desarrollo de escalas. Con fines ilustrativos se expone el análisis de cuatro ejemplos numéricos y se proporcionan los datos y la sintaxis en R.
Palabras clave: Fiabilidad; consistencia interna; coeficiente alfa; coeficiente omega; medidas congenéricas; medidas tau-equivalentes; análisis factorial confirmatorio.

Title: A journey around alpha and omega to estimate internal consistency reliability.
Abstract: Based on recent psychometric developments, this paper presents a conceptual and practical guide for estimating internal consistency reliability of measures obtained as item sum or mean. The internal consistency reliability coefficient is presented as a by-product of the measurement model underlying the item responses. A three-step procedure is proposed for its estimation, including descriptive data analysis, test of relevant measurement models, and computation of internal consistency coefficient and its confidence interval. Provided formulas include: (a) Cronbach's alpha and omega coefficients for unidimensional measures with quantitative item response scales, (b) coefficients ordinal omega, ordinal alpha and nonlinear reliability for unidimensional measures with dichotomic and ordinal items, (c) coefficients omega and omega hierarchical for essentially unidimensional scales presenting method effects. The procedure is generalized to weighted sum measures, multidimensional scales, complex designs with multilevel and/or missing data and to scale development. Four illustrative numerical examples are fully explained and the data and the R syntax are provided.
Key words: Reliability, internal consistency, coefficient alpha, coefficient omega, congeneric measures, tau-equivalent measures, confirmatory factor analysis.

Hubo una época en que el coeficiente alfa de Cronbach (α, Cronbach, 1951) era ampliamente aceptado como indicador de la fiabilidad de un cuestionario diseñado para medir un único constructo. Era el estimador de la fiabilidad de consistencia interna de la suma o promedio de las respuestas a los ítems. Bajo el paraguas de la teoría clásica de los tests (TCT, Lord y Novick, 1968), el cálculo de α se aplicaba a ítems con escala de respuesta cuantitativa, de la misma manera que sus formulaciones equivalentes se aplicaban a ítems dicotómicos, como es el caso de la expresión KR-20, o a respuestas estandarizadas, en el caso de la fórmula de Spearman-Brown (e.g., Muñiz, 1992; Nunnally, 1978).

No era relevante que el primer autor que publicó la formulación del coeficiente no fuera Cronbach (e.g., Revelle y Zinbarg, 2009), ni que el propio Cronbach alertase contra su uso abusivo (Cronbach y Shavelson, 2004), ni tampoco las numerosas llamadas a su sustitución, ampliamente argumentadas por un numeroso grupo de psicómetras (Bentler, 2009; McDonald, 1999; Raykov, 1997; Zinbarg, Revelle, Yovel, y Li, 2005). Tampoco importaba que la medida analizada fuera una suma ponderada de ítems, como ocurre en los modelos de ecuaciones estructurales con variables latentes (SEM). En todos los casos, el cálculo de α era previo a cualquier análisis relacionado con un constructo. Su papel cumplía con la necesidad, indicada en las directrices del manual de publicaciones de la American Psychological Association (2010), de informar acerca de la calidad psicométrica de las medidas y covariables utilizadas.

El éxito de α y su supervivencia en la literatura científica es atribuible a muchos motivos. Se aplica a una forma sencilla y estable de medir un constructo mediante la simple suma o promedio de respuestas a los ítems; se puede compartir fácilmente con los revisores y lectores de artículos del ámbito de las Ciencias Sociales y de la Salud; puede obtenerse con un diseño sencillo, basado en una sola administración del cuestionario; se calcula con facilidad gracias a la ayuda de diversos paquetes o entornos de software estadístico como SPSS, SAS o Stata. De esta forma, el uso de α se ha convertido en un nuevo ejemplo del conocido divorcio entre las publicaciones metodológicas y las aplicadas en psicología durante los primeros años del siglo XXI. Otros ejemplos del mencionado divorcio pueden verse en Izquierdo, Olea y Abad (2014) o Lloret-Segura, Ferreres-Traver, Hernández-Baeza y Tomás-Marco (2014).

Revisando la literatura psicométrica del siglo XXI sobre el uso de α, nos ha venido a la memoria la épica circunnavegación del globo hecha en el siglo XVI por marineros capitaneados por Magallanes y Elcano. La expedición partió de Sanlúcar de Barrameda y después de tres años de peligrosa navegación hacia Occidente, regresó habiendo completado un viaje alrededor de nuestro planeta. Cuando la nao Victoria alcanzó de nuevo el lugar de partida, el conocimiento que se había adquirido durante el viaje condicionaría definitivamente el futuro. *Mutatis mutandis*, durante los últimos años se ha realizado un gran esfuerzo en el ámbito psicométrico por proporcionar indicadores de fiabilidad de consistencia interna alternativos a α. Estos coeficientes alternativos, por lo general, están basados en el modelo de medida subyacente a cada cuestionario y en los estimadores apropiados para cada tipo de datos. Después de años discutiendo esos nue-

* **Correspondence address [Dirección para correspondencia]:** Carme Viladrich. Departament de Psicobiologia i Metodologia de les Ciències de la Salut. C. de la Fortuna s/n. 08193 Bellaterra, Cerdanyola del V. (España). E-mail: carme.viladrich@uab.cat

vos indicadores, la Psicometría parece haber regresado al punto de partida. En este sentido puede verse, por ejemplo, la viva discusión sostenida entre partidarios de los indicadores clásicos y de los nuevos indicadores en la revista *Educational Measurement. Issues and Practice* (Davenport, Davison, Liou, y Love, 2016 y sus referencias). Más relevantes son las recientes publicaciones que explícitamente sugieren una vuelta al uso de α siempre y cuando pueda mostrarse que este coeficiente proporciona una estimación correcta de la fiabilidad (Green, et al, 2016; Raykov, West y Traynor, 2015). La consecuencia más importante de esta particular vuelta al mundo de la Psicometría es que actualmente ya no se puede calcular la fiabilidad de consistencia interna utilizando α de forma inocente haciendo unos pocos clics en el menú de un *software* estadístico. Tal vez α sea adecuado para un determinado tipo de datos, pero habrá que argumentarlo con la comprobación de algunos supuestos (ver apartado siguiente). Si estos supuestos no se cumplen, deberían utilizarse coeficientes alternativos derivados del modelo de medida. Durante este largo viaje, la fiabilidad de consistencia interna ha pasado de ocupar una posición central como concepto psicométrico a ser un subproducto de un modelo de medida; algo que, por otra parte, no resulta una novedad para las personas familiarizadas con los modelos de medida psicométricos (Birnbaum, 1968; Jöreskog, 1971), pero que no ha sido incluido de forma rutinaria en el ámbito aplicado del desarrollo y evaluación de escalas. Aquellos estimadores adimensionales (*dimension free* en inglés), no basados en un modelo de medida específico, como la β de Revelle o el *greatest lower bound*, siguen siendo objeto de controversia (Bentler, 2009; Raykov, 2012; Revelle y Zinbarg, 2009; Sijtsma, 2009, 2015) y no serán considerados en este artículo.

Afortunadamente, los indicadores basados en modelos de medida, aunque por lo general requieren de una muestra grande, sólo necesitan una administración del cuestionario y son fáciles de calcular gracias a mayor accesibilidad del *software* actual. Entre las opciones más usuales se encuentran el entorno y lenguaje de programación libre R (R Core Team, 2016) y el programa comercial Mplus (Muthén y Muthén, 2015). Así las cosas, creemos que el siguiente paso consiste en facilitar, tanto a los autores como a los revisores, la incorporación rutinaria de este conocimiento en su trabajo para mejorar la calidad de las publicaciones que incluyen medidas basadas en cuestionarios del ámbito de las Ciencias Sociales y de la Salud.

Nuestra voz se añade a la de otros autores como Brunner, Nagy, y Wilheim (2012), Crutzen y Peters (2015), Graham (2006) o Green y Yang (2015). En comparación con ellos, nuestro trabajo tiene un enfoque más procedimental e incluye varias contribuciones específicas, a saber: la necesidad explícita de incluir en el procedimiento la exploración previa de los datos y la manera de llevarla a cabo; un esquema de los métodos de estimación y de los índices de bondad de ajuste para los modelos SEM con variables cuantitativas y ordinales; un conjunto completo de fórmulas y procedimientos para la estimación puntual y por intervalo de confianza (IC) de la fiabilidad de consistencia interna; un método para determinar en qué momento α sería prácticamente comparable a los índices basados en SEM; una forma práctica de llevar a cabo el análisis completo en R; y finalmente, un diagrama de decisiones a modo de síntesis del análisis.

El objetivo de este trabajo es proporcionar un conjunto actualizado de reglas prácticas para estudiar la fiabilidad de consistencia interna de la suma o promedio de las respuestas a los ítems diseñados para medir un solo constructo, una forma de puntuación compuesta en la que todos los ítems tienen el mismo peso. Se proporciona una justificación para el uso de dichas reglas, así como ejemplos de aplicación de las mismas a diferentes tipos de datos. Por otra parte se proporcionan apéndices con la sintaxis en R comentada. Estos apéndices van dirigidos tanto a los investigadores experimentados con el uso de R como a aquellos que no están familiarizados con este *software*. Además, se discute su generalización a modelos de medida complejos, así como sus consecuencias prácticas para el diseño y el análisis de los datos.

En lo que sigue, este artículo se estructura en cinco apartados. En primer lugar, se exponen los conceptos básicos de modelo de medida y de fiabilidad. En este contexto, se presentan los casos donde el uso de α puede ser adecuado. A continuación, se incluye la aplicación práctica de estos conceptos en tres fases: (a) exploración de los datos, (b) ajuste del modelo de medida, y (c) estimación de la fiabilidad de consistencia interna. El procedimiento es aplicado a ítems con escala de respuesta cuantitativa y a ítems con escala de respuesta ordinal. Seguidamente, se resuelven por completo cuatro casos que ilustran su uso en escenarios habituales en la investigación aplicada. En el cuarto apartado se discute la aplicación de este procedimiento a situaciones más complejas como son los modelos multidimensionales, los ítems que pesan en más de un factor, los diseños con datos faltantes y/o multinivel y el desarrollo de nuevas escalas. Concluimos el artículo con un resumen práctico de las principales recomendaciones, incluido un diagrama de toma de decisiones.

Modelo de medida y coeficiente de fiabilidad de las puntuaciones compuestas unidimensionales

De acuerdo con la TCT, las respuestas observadas son el resultado de sumar una puntuación verdadera (*T*) o sistemática y un término de error aleatorio (*E*), que tiene media cero y no tiene relación con el resto de las variables. El coeficiente de fiabilidad es definido como la razón de la varianza de la puntuación verdadera y la varianza de la puntuación observada, que a su vez es la suma de la puntuación verdadera más la varianza del error (Lord y Novick, 1968):

$$\rho = \frac{\sigma^2(T)}{\sigma^2(T) + \sigma^2(E)} \qquad (1)$$

Los valores del coeficiente de fiabilidad están comprendidos entre 0 y 1 y, en general, se consideran aceptables los valores superiores a .7 cuando se está desarrollando una nueva medida, los valores superiores a .8 cuando se aplican a investigación y, por ejemplo, se comparan medias grupales, y los valores superiores a .90 cuando las puntuaciones se utilizan para tomar decisiones importantes que afectan individualmente a personas (Nunnally, 1978). Una propiedad bien conocida del coeficiente es que la varianza verdadera no depende únicamente de las características del cuestionario sino también de la variabilidad del constructo en la población analizada. Manteniendo constantes todos los

demás aspectos, cuanto más variable sea el constructo mayor será la fiabilidad.

Puesto que la TCT es un modelo meramente teórico, son necesarias estrategias para obtener la estimación empírica del coeficiente de fiabilidad y la más habitual es la de la consistencia interna basada en un diseño que requiere una única administración de una sola prueba. Este enfoque implica el supuesto adicional de que las respuestas a los ítems comparten un único constructo subyacente y permite que las varianzas verdaderas y totales se deriven de las estimaciones de los parámetros de un modelo de análisis factorial confirmatorio (AFC; e.g., McDonald, 1999). El modelo de medida subyacente a un AFC está representado gráficamente en la Figura 1 donde, por convención, cada ítem (Y_j) está representado en cuadrados puesto que se trata de variables observables, el constructo o factor (F) y los errores (ε_j) se representan en óvalos, puesto que no son variables directamente observables, y las relaciones entre variables están representadas por flechas.

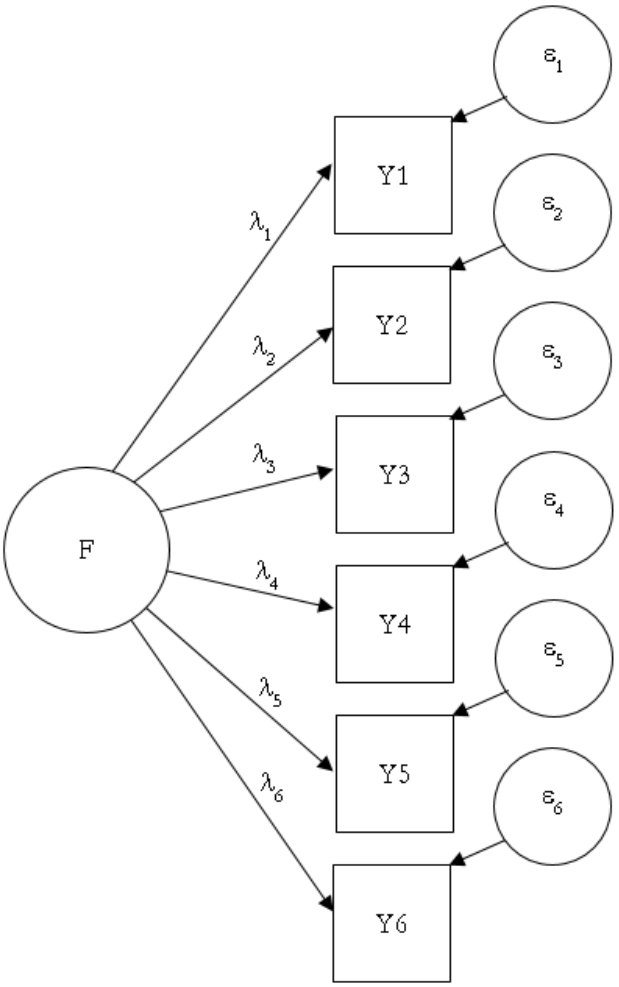

Figura 1. Modelo de medida con seis ítems pesando en un único factor.

El constructo es una variable latente, es decir, no observable directamente sino inferida a partir de las respuestas a los ítems que son variables observables. La relación entre el constructo y el ítem es lineal y está cuantificada por la carga factorial (λ_j). Lambda es una medida de la discriminación del ítem que se interpreta como un coeficiente de regresión: cuando el valor del factor incrementa en una unidad, el valor del ítem j incrementa en λ_j unidades. Conviene señalar que la linealidad únicamente es apropiada para ítems con distribuciones normales. Cada ítem también está caracterizado por su índice de dificultad, cuantificado en AFC por la constante o puntuación del ítem cuando el factor toma el valor cero. Finalmente, el término de error es único para cada ítem, está incorrelacionado con la puntuación factorial y también con los errores de los otros ítems.

Estableciendo la varianza del factor a 1 con la finalidad de identificar el modelo, se puede demostrar (Jöreskog, 1971; McDonald, 1999) que la fiabilidad de una puntuación obtenida como suma o promedio de los ítems es:

$$\omega = \frac{(\sum \lambda_j)^2}{(\sum \lambda_j)^2 + \sum \sigma^2_{\varepsilon_j}} \qquad (2)$$

o la razón entre la varianza de la puntuación verdadera, derivada de la estimación de los parámetros del modelo, y la suma de las varianzas y covarianzas reproducidas por el modelo. Este estimador ha recibido varias denominaciones; coeficiente omega por McDonald, fiabilidad compuesta por Raykov (1997), y puesto que el AFC es una parte de los procedimientos SEM, fiabilidad de consistencia interna estimada por SEM por otros autores (e.g., Yang y Green, 2011). En Raykov (2012) se puede encontrar una ecuación más general basada en una variable latente no estandarizada.

De acuerdo con McDonald (1999), si el modelo de medida se ajusta a los datos, la Ecuación 2 puede ser reescrita sustituyendo el denominador por la suma de las varianzas observadas de los ítems (σ_j^2) y las covarianzas entre ítems ($\sigma_{j<j'}$):

$$\omega = \frac{(\sum \lambda_j)^2}{\sum \sigma_j^2 + 2\sum \sigma_{j<j'}} \qquad (3)$$

De hecho, McDonald considera que la Ecuación 3 es más conveniente que la Ecuación 2. Según otros expertos, como Bentler (2009), la matriz de covarianzas reproducida por un modelo es una estimación más eficiente de la matriz de covarianzas poblacional que la estimación producto-momento. En caso de que el modelo se ajuste a los datos, las consecuencias prácticas serán inapreciables. Por el contrario, en caso que el modelo de medida no se ajuste a los datos, compartimos la recomendación de McDonald de que no debería utilizarse ninguna de las expresiones del coeficiente omega para estimar la fiabilidad de consistencia interna.

Como veremos a continuación, actualmente omega hace referencia a una familia de coeficientes de fiabilidad de consistencia interna derivados de las estimaciones de los parámetros AFC. La mayoría de estos coeficientes proceden de la relajación de los supuestos de errores no correlacionados, normalidad y unidimensionalidad con el fin de adaptarlos a las propiedades de los datos reales. El propio coeficiente alfa no es sino un miembro de la misma familia basado en supuestos muy restrictivos.

Fiabilidad de las medidas esencialmente tau-equivalentes

El omnipresente α es un estimador insesgado de la fiabilidad de consistencia interna siempre que el modelo de medidas esencialmente tau-equivalentes se ajuste a los datos (Jöreskog, 1971; McDonald, 1999). Este modelo se representa a la izquierda de la Figura 2. Obsérvese que las cargas de todos los ítems en el factor han sido igualadas. Esto refleja el supuesto de que todos los parámetros de discriminación son iguales, es decir, que si se controla la diferencia de puntuaciones factoriales entre dos grupos de examinados, la diferencia de las puntuaciones de los ítems entre los dos grupos será constante en todos los ítems.

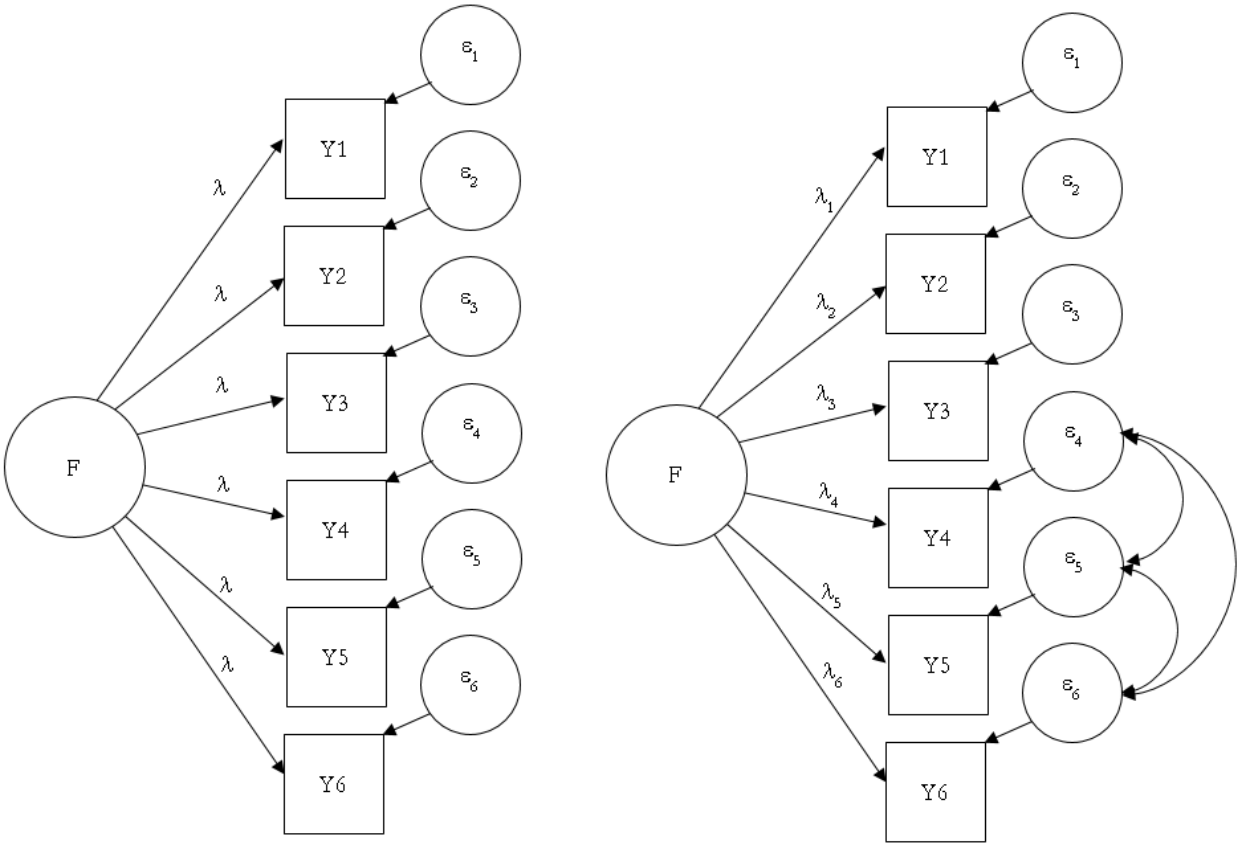

Figura 2. Modelo de medidas esencialmente tau-equivalentes a la izquierda, y modelo de medidas con errores correlacionados a la derecha.

Si la tau-equivalencia esencial se mantiene, el valor del coeficiente omega es igual al valor de α que, a su vez, es igual a otros coeficientes desarrollados anteriormente para los mismos propósitos, como la Lambda3 de Guttmann (e.g., Revelle y Zinbarg, 2009). El numerador de la Ecuación 3 anterior se reduce en este caso al producto del número de ítems al cuadrado (k^2) por la carga del factor al cuadrado (λ^2). El estimador de mínimos cuadrados no ponderados (ULS) de las cargas factoriales al cuadrado es el promedio de covarianzas entre los ítems, y el denominador es la suma de las varianzas y covarianzas observadas entre los ítems.

$$\omega = \alpha = \frac{k^2\lambda^2}{\sum \sigma_j^2 + 2\sum \sigma_{j<j'}} \qquad (4)$$

Consecuentemente, si el modelo de medidas esencialmente tau-equivalentes se ajusta a los datos, se podría calcular α para proporcionar una estimación de la fiabilidad de consistencia interna de la suma o promedio de los ítems.

Fiabilidad de las medidas congenéricas

Como cualquier persona con experiencia en AFC sabe, las cargas factoriales de los ítems no suelen ser iguales en un análisis visual. Esta desigualdad queda mejor especificada por el modelo de medidas congenéricas, que permite diferentes valores de discriminación entre los ítems. De hecho, este es el modelo factorial unidimensional general representado en la Figura 1 que se ha explicado con anterioridad. Si el modelo de medidas congenéricas se ajusta a los datos y el modelo de medidas esencialmente tau-equivalentes, más restrictivo, no lo hace, la consistencia interna de la suma o promedio de los ítems debe estimarse mediante el coeficiente omega utilizando la Ecuación 2 o la Ecuación 3.

La relación entre α y omega para medidas congenéricas no esencialmente tau-equivalentes ha sido estudiada a fondo. En primer lugar, se ha demostrado que en este caso α es menor que omega y, por lo tanto, se puede utilizar como límite inferior de la fiabilidad (Raykov, 1997). En segundo lugar, los estudios de simulación han demostrado que la diferencia entre α y omega no tiene consecuencias prácticas cuando las cargas factoriales son, en promedio, .70 y las diferencias entre ellas están dentro del intervalo -.20 y +.20 (Raykov y Marcoulides, 2015). Por lo tanto, si se cumplen estas condiciones, podemos seguir utilizando α como el estimador puntual de la fiabilidad de consistencia interna, algo que, según estos autores, puede ser incluso deseable por razones prácticas. De lo contrario, se debería usar omega puesto que α subestimaría la fiabilidad de consistencia interna, al menos en el caso de encontrarse una diferencia estadísticamente significativa entre α y omega (Deng y Chan, 2016).

Finalmente, los estudios de simulación (Gu, Little y Kingston, 2013), han demostrado que ni el número de ítems no tau-equivalentes de un cuestionario, ni la magnitud de las diferencias entre las cargas factoriales producen sesgos importantes al usar α para estimar la fiabilidad poblacional. Los sesgos más grandes se deben a la presencia de errores correlacionados y a razones pequeñas entre la varianza verdadera y la varianza del error.

Fiabilidad de las medidas con errores correlacionados

Trataremos ahora las medidas en las que el supuesto de errores independientes no se sostiene. Un caso bien conocido ocurre cuando un cuestionario contiene ítems que miden el mismo constructo redactados de forma directa e inversa. En este caso, una vez que se controla el efecto de la variable latente, los ítems con expresión positiva conservan una covarianza no despreciable entre sí, al igual que los ítems redactados negativamente. Esta situación puede modelizarse especificando algunas correlaciones distintas de cero entre errores (Figura 2, derecha; e.g., Brown, 2015; Marsh, 1996), o bien como un factor de método debido a la composición del cuestionario (Figura 4; ver más adelante y también Gu et al., 2013) o incluso como un parámetro debido a las diferencias individuales de los encuestados (Maydeu-Olivares y Coffman, 2006). En aras de la simplicidad, en este apartado nos centraremos en la primera de las opciones enunciadas y nos referiremos brevemente al resto al tratar con la suposición de unidimensionalidad.

Si no se tiene en cuenta, la presencia de errores correlacionados tiene efectos graves sobre la estimación de la fiabilidad de consistencia interna. En este caso, las estimaciones de las cargas factoriales serían incorrectas (e.g., Brown, 2015) y tanto omega como α serían estimadores de la fiabilidad poblacional sesgados, aunque el sesgo sería mucho mayor si se utilizase α (Gu et al., 2013). Además, α ya no podría ser considerado como el límite inferior de fiabilidad de las puntuaciones (Raykov, 2001). De hecho, dependiendo de la configuración de los parámetros del modelo de medida, el sesgo de α podría llevar a subestimar, o incluso peor, a sobrestimar la fiabilidad en la población, dando la falsa sensación de que las puntuaciones son fiables cuando realmente no lo serían.

Este sesgo debe ser corregido incluyendo la covarianza entre los errores tanto en la estimación de los parámetros del modelo como en la fórmula para el cálculo de omega tal y como se muestra a continuación para omega (Raykov, 2004, véase Bollen, 1980 para la formulación inicial con el factor no estandarizado):

$$\omega = \frac{(\sum \lambda_j)^2}{(\sum \lambda_j)^2 + \sum \sigma^2_{\varepsilon_j} + 2\sum \sigma_{\varepsilon_{j<j'}}} \qquad (5)$$

La suma de los elementos de la matriz de varianzas-covarianzas reproducida por el modelo en el denominador ilustra claramente la diferencia entre la Ecuación 5 y la Ecuación 2. De nuevo, si el modelo con errores correlacionados se ajusta a los datos y, por tanto, los parámetros del modelo se han estimado correctamente, se obtendrían resultados muy similares utilizando en el denominador de la Ecuación 3 la matriz de varianzas-covarianzas observada.

Los estudios de simulación de Gu et al. (2013) demostraron que, en presencia de errores correlacionados, α puede sobrestimar la fiabilidad en la población con diferenciales tan altos como .38. Esto significaría, por ejemplo, que una puntuación con una fiabilidad real de .40, que es completamente inaceptable, puede dar como resultado un valor de α de hasta .78, lo que puede conducir a la conclusión errónea de que se trata de una fiabilidad bastante buena. Estos autores muestran también que manteniendo iguales todas las condiciones, el coeficiente omega corregido con la correlación entre errores tendría un sesgo de -.09, prácticamente despreciable y por lo tanto preferible. Su conclusión es que α parece tratar la correlación entre errores como si fuera parte de la varianza verdadera produciendo así una sobreestimación de la fiabilidad. Volveremos a este punto a más adelante cuando tratemos la hipótesis de la unidimensionalidad.

Por qué y cómo desarrollar el análisis

La estimación incorrecta de la fiabilidad tiene consecuencias indeseables en todos los campos aplicados en donde se utilizan cuestionarios. En el desarrollo de instrumentos, la subestimación de la fiabilidad puede llevar a que los investigadores intenten mejorar el cuestionario innecesariamente, mientras que su sobreestimación produciría una injustificada confianza en el mismo. Incluso si en el momento de desarrollar el cuestionario pudiera ser suficiente obtener el límite inferior de la fiabilidad de sus puntuaciones, éste no bastaría para su uso posterior en otros contextos. En investigación básica o aplicada, si se utilizase esa estimación sesgada de la fiabilidad para calcular la corrección por atenuación, los tamaños del efecto podrían verse seriamente afectados (Revelle y Zinbarg, 2009). En la toma de decisiones individuales, una estimación de fiabilidad sesgada afectaría el error estándar de medida, lo que podría llevar a decisiones inadecuadas en la interpretación y comunicación de las puntuaciones.

La mejor manera de asegurar una estimación correcta de la fiabilidad basada en un diseño de consistencia interna, consiste en derivarla de un modelo de medida ajustado utilizando métodos SEM. En este apartado presentamos un procedimiento ba-

sado en tres fases analíticas posponiendo la discusión de los costes de esta estrategia de análisis hasta el último apartado del artículo.

Fase 1: Exploración de las respuestas a los ítems. Se estudiarán las distribuciones univariadas y las relaciones entre ítems. Con ello se tomarán decisiones sobre los tipos de variables y las posibles agrupaciones de ítems que puedan afectar a la especificación y estimación del modelo en la siguiente fase.

Fase 2: Ajuste del modelo de medida a los datos. Se trata de una actividad confirmatoria, y por tanto, comienza con la especificación de los modelos derivados del conocimiento previo del cuestionario, continúa con la estimación de los parámetros y finaliza con la evaluación tanto de la bondad de ajuste y como de la adecuación de la solución en cada caso. El objetivo es elegir el modelo de medida que tenga un significado conceptual y un buen ajuste a los datos. Para los cuestionarios que son supuestamente unidimensionales, el analista deberá considerar los modelos de medidas esencialmente tau-equivalentes, congenéricas y, quizás, el de errores correlacionados. Estos modelos están anidados, siendo el modelo de medidas esencialmente tau-equivalentes el más más restrictivo (es decir, con menos parámetros libres a estimar) y el modelo de errores correlacionados el menos restrictivo.

Fase 3: Cálculo del coeficiente de fiabilidad derivado de los parámetros del modelo de medida elegido en la fase anterior y de su error estándar de medida para realizar la estimación por intervalo.

Nuestra propuesta está en la línea de las de otros autores que sugieren realizar siempre la Fase 2 para derivar la estimación de la fiabilidad a la Fase 3 del análisis (e.g., Crutzen y Peters, 2015; Graham, 2006; Green y Yang, 2015). Para mayor claridad, además de estas dos fases consensuadas, consideramos esencial enfatizar otra etapa implícita en el análisis. Se debe realizar una exploración previa de los datos para decidir correctamente tanto la matriz de asociaciones a analizar como el estimador de los parámetros del modelo de medida. En este sentido, véanse, por ejemplo, los capítulos de Behrens, DiCerbo, Yel y Levy (2012), Malone y Lubansky (2012) y Raykov (2012) en sendos libros de texto, o los artículos de Lloret-Segura et al. (2014) y de Ferrando y Lorenzo-Seva (2014) en un número anterior de esta revista. En pocas palabras, además de tener en cuenta el tamaño de la muestra y la completitud de los datos, en la Fase 1 del análisis deben analizarse las distribuciones de las respuestas a los ítems y las relaciones entre ellas para determinar el tipo de datos y detectar posibles ítems relacionados diferencialmente con el resto o incluso agrupaciones de ítems. Esta fase permitirá a quien analiza decidir si tratar sus datos como cuantitativos o como ordinales, dos opciones que se abordarán en los dos apartados siguientes, y también decidir sobre una posible corrección por la presencia de errores correlacionados, cuyo tratamiento se verá con más detalle en el apartado que contiene los escenarios prácticos.

Además, en la Fase 3 sugerimos la estimación por intervalo de la fiabilidad. Aunque se acostumbra a publicar la estimación puntual de los coeficientes de consistencia interna, debe tenerse en cuenta que éstos son indicadores estadísticos obtenidos en una muestra y, por lo tanto, afectados por el error estándar. En consecuencia, sugerimos la publicación del IC con una confianza del 95% tal como es costumbre en las Ciencias Sociales y de la Salud. El error estándar para los coeficientes de fiabilidad puede estimarse mediante *bootstrap* (Kelley y Pornprasertmanit, 2016; Raykov y Marcoulides, 2015) o aproximarse utilizando el método analítico delta (Raykov, 2012; Padilla y Divers, 2016). El método delta es menos costoso desde un punto de vista computacional y proporciona resultados comparables al *bootstrap* con ítems no binarios y muestras grandes (más de 250 casos, Padilla y Divers, 2016).

Análisis de datos cuantitativos

Las respuestas a los ítems obtenidas con una escala cuantitativa (e.g., escala analógica visual) son datos cuantitativos continuos. Las respuestas obtenidas en una escala como las de tipo Likert, son datos ordinales que pueden ser analizados como variables continuas siempre que el número de categorías sea alto (5 o más) y la distribución de frecuencia no muestre efectos suelo o techo (Rhemtulla, Brosseau-Liard y Savalei, 2012). Esta es la decisión principal que se tomará en la Fase 1.

En la Fase 2, se especificarán con AFC todos los modelos de medida que resulten plausibles para los datos disponibles. Por lo menos, deberían considerarse el modelo de medidas esencialmente tau-equivalentes y el de medidas congenéricas. A continuación, se estimarán los parámetros del modelo, se calcularán los índices de bondad de ajuste y se elegirá el modelo de medición con mejor ajuste que a la vez sea parsimonioso e interpretable. En la Fase 3 del análisis se utilizarán los parámetros estimados para obtener la fiabilidad de consistencia interna. Todas estas operaciones se pueden realizar con bastante facilidad utilizando *software* comercial como Mplus (Muthén y Muthén, 2017) y también el entorno de *software* libre R (R Core Team, 2016). En apartados posteriores se presentan ejemplos de sintaxis en R.

Describimos aquí los procedimientos para el ajuste del modelo SEM, pero los detalles completos exceden los objetivos de este documento. Para un tratamiento en profundidad sobre la estimación de parámetros, ajuste de modelos, comparación y revisión de modelos, véanse referencias como Abad, Olea, Ponsoda y García (2011), Brown (2015) o Hoyle (2012).

Se analizará la matriz de datos completa de casos por ítems o la matriz de varianzas-covarianzas entre ítems. Si la distribución normal multivariable es aceptable, se utilizará el método de estimación de máxima verosimilitud (ML) y se evaluará la bondad del ajuste utilizando indicadores de ajuste global y local. Un valor de χ^2 estadísticamente nulo, junto con los valores de los parámetros y los errores estándar dentro de un rango aceptable, proporcionarían evidencia favorable al modelo de medida que se está probando. Complementariamente, también pueden tomarse decisiones utilizando índices de ajuste aproximados, tales como el índice de ajuste comparativo (CFI), el índice de Tucker-Lewis (TLI) y el error cuadrático medio de aproximación (RMSEA), todos ellos con valores entre 0 y 1. En líneas generales, para que el modelo se considere apropiado, los valores de CFI y TLI deben ser mayores que .95 y los de RMSEA inferiores a .05.

Los modelos anidados pueden compararse con base en la diferencia de sus valores χ^2 asociados. Esta comparación formal también puede complementarse evaluando las diferencias entre los índices de ajuste aproximados. En general, se considera que

dos modelos anidados ajustan igualmente bien a los datos si la diferencia de χ^2 es estadísticamente no significativa y también si las diferencias entre los índices aproximados de ajuste son inferiores a .01.

En el caso de datos que presenten desviaciones menores de la normalidad, incluso si se trata de categorías ordenadas sin efectos de suelo o techo, se puede utilizar la estimación de máxima verosimilitud robusta (MLR) y los índices χ^2, CFI, TLI y RMSEA asociados. En este caso, los parámetros se estiman igualmente por ML, pero sus errores estándar y los índices de ajuste global se calculan corregidos respecto a la no-normalidad. La comparación entre los modelos anidados, sin embargo, no es tan directa, ya que la diferencia entre los valores χ^2 corregidos no es interpretable. Se deben aplicar los factores de corrección de Satorra-Bentler o de Yuan-Bentler (Muthén y Muthén, s.f.).

Independientemente del estimador utilizado, la Fase 2 del análisis concluye con la elección del modelo más parsimonioso, con sentido conceptual y un buen ajuste a los datos. Los parámetros estimados durante la Fase 3 del análisis se utilizarán, en los casos apropiados, para calcular los coeficientes alfa u omega, mientras que el error estándar para la estimación por intervalo se obtendrá generalmente con *bootstrap*, o si se dispone de muestras grandes, mediante el método delta.

Análisis de datos ordinales y dicotómicos

Muchos cuestionarios tienen formatos de respuesta categóricos con dos opciones (e.g., Sí / No) o más (e.g., Muy en desacuerdo / En desacuerdo / De acuerdo / Muy de acuerdo). Por lo tanto, a menudo el analista se enfrenta a datos categóricos binarios u ordinales. En la Fase 1 del análisis, se prestará especial atención al número de categorías de respuesta que realmente han sido utilizadas por las personas encuestadas y también a la forma de su distribución. Si el número de categorías de respuesta es de cuatro o menos, o también de cinco o más con importantes efectos suelo o techo, la estimación de parámetros que se realice en la siguiente fase ya no podrá ser aproximada por estimadores basados en la normalidad. En estos casos debería utilizarse una estrategia apropiada para datos categóricos, siempre de acuerdo con la recomendación de Rhemtulla, Brosseau-Liard, y Savalei (2012) que también hemos utilizado en el apartado anterior.

Para desarrollar la Fase 2 del análisis, se podrá elegir entre tres opciones (e.g., Bovaird y Koziol, 2012). La primera consiste en, antes de realizar el análisis, agrupar varios ítems (*parceling* en inglés) lo que daría lugar a datos cuantitativos. Esta solución sigue siendo muy controvertida (Little, Rhemtulla, Gibson, y Schoemann, 2013; Marsh, Lüdtke, Nagengast, Morin, y Von Davier, 2013) y sólo es creíble si resulta estable entre diferentes formas igualmente plausibles de parcelar los ítems (Raykov, 2012). La segunda opción consiste en no agrupar los ítems y estimar los parámetros de cada uno de ellos con un modelo de teoría de respuesta al ítem (TRI) que resulte plausible para esos datos. Esta estrategia utilizaría la estimación con información completa (basada en los patrones de respuesta) y se podría aplicar a modelos de uno, dos, tres o cuatro parámetros. La tercera opción sigue siendo un análisis a nivel de ítem, es aplicable únicamente a los modelos de uno y dos parámetros, y utiliza la estimación con información limitada (basada en la matriz de correlaciones policóricas o tetracóricas) típica de los modelos AFC. En este artículo adoptamos esta tercera opción ya que facilita la generalización de los conceptos tratados hasta ahora y se ha demostrado su equivalencia con los modelos TRI de ojiva normal más usuales (e.g., Cheng, Yuan y Liu, 2012; Ferrando y Lorenzo-Seva, 2017).

El modelo para respuestas a ítems ordinales está representado en la Figura 3. Para dar cuenta del carácter ordinal de los ítems, se define una distribución de respuesta continua latente (Y_j^*) que determina una distribución de categorías ordenadas observadas (Y_j). La respuesta latente está relacionada con la respuesta observada a través de umbrales discretos (Muthén, 1984). En otras palabras, cuando hay un cambio en el valor de Y_j^* que cruza el umbral entre dos categorías de respuesta, el valor discreto observado en Y_j cambia a la categoría adyacente. Generalmente, se toma la distribución normal acumulada como función de enlace entre los umbrales y las proporciones acumuladas de respuestas. Por lo demás, el modelo latente para Y_j^* sería el mismo que el modelo de la Figura 1. Por lo tanto, los modelos de medida a considerar seguirán siendo los de medidas esencialmente tau-equivalentes, de medidas congenéricas y el de medidas con errores correlacionados. Los cambios se producen únicamente en las técnicas de estimación. Hoy en día, la mayoría de los paquetes estadísticos para SEM incluyen opciones para ajustar correctamente modelos de medición para datos ordinales. También en este caso el *software* comercial Mplus y el entorno de *software* libre R son de los más usados. En posteriores apartados se muestra un ejemplo de tratamiento de datos ordinales con la sintaxis correspondiente desarrollada en R.

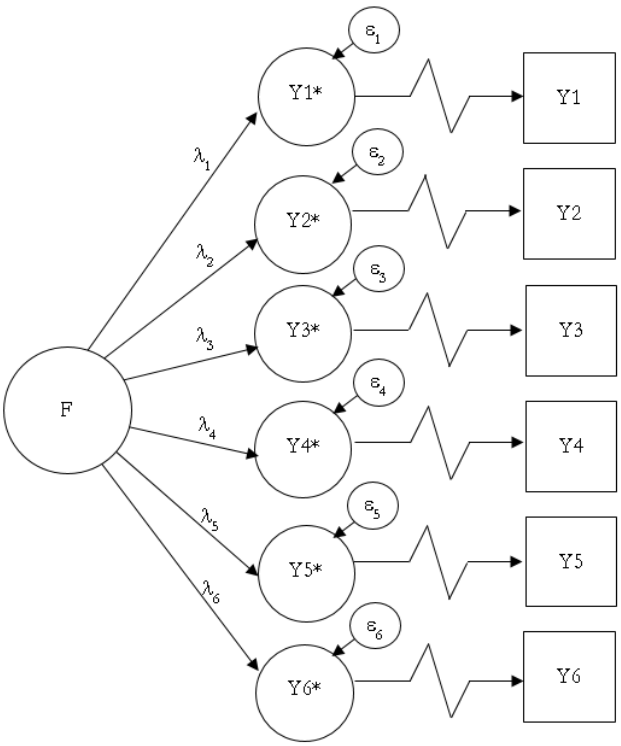

Figura 3. Modelo de medida con seis ítems contestados en una escala de respuesta ordinal pesando en un único factor.

Al igual que hicimos con los datos cuantitativos, a continuación describimos el esquema del procedimiento para la Fase 2. Para un tratamiento más en profundidad, ver por ejemplo, Brown (2015), Hancock y Mueller (2013), o Hoyle (2012). El procedimiento comienza con la estimación de la matriz de correlaciones policóricas para ítems con tres o más categorías, o de correlaciones tetracóricas para ítems dicotómicos. En segundo lugar, se ajusta el modelo de medida a esta matriz de correlaciones utilizando un método de estimación adecuado a la naturaleza categórica de las variables. El método de estimación más adecuado para una amplia gama de tamaños de muestra es el de mínimos cuadrados ponderados robusto con un estadístico χ^2 ajustado por media y varianza (WLSMV; e.g., Bovaird y Koziol, 2012), aunque para muestras pequeñas (alrededor de 200 casos) el método ULS puede ser una buena alternativa (Forero, Maydeu-Olivares, y Gallardo-Pujol, 2009). En este caso, la interpretación de los resultados, incluida la de los índices de bondad de ajuste y la comparación de los modelos anidados, requieren todavía más formación por parte del analista, por lo que recomendamos consultar los textos especializados mencionados y estar al día sobre los nuevos desarrollos en este campo (e.g., Huggins-Manley, y Han, 2017; Maydeu-Olivares, Fairchild, y Hall, 2017; Sass, Schmitt, y Marsh, 2014).

Una vez obtenidas las estimaciones de los parámetros del modelo, se puede estimar la consistencia interna de la suma de los ítems latentes Y_j^* a partir del coeficiente omega ordinal (Elosua y Zumbo, 2008; Gadermann, Guhn y Zumbo, 2012). Estos autores proponen calcular el coeficiente, de acuerdo con la Ecuación 2, a partir de los parámetros estimados, tanto en el numerador, donde se obtiene la varianza de la puntuación verdadera a partir de las cargas factoriales estimadas, como en el denominador, donde la varianza de la puntuación verdadera más la de la varianza del error se obtiene por la suma de todos los elementos de la matriz de correlaciones policóricas. Como es habitual, si el modelo se ajusta bien a los datos, se pueden utilizar los elementos de la matriz de correlaciones policóricas en coherencia con la Ecuación 3. Por otra parte, el denominador de la ecuación puede simplificarse sumando los elementos de la matriz de correlaciones, por lo que para un cuestionario con k medidas ordinales congenéricas, omega se reduce a la Ecuación 6, donde ρ^* se refiere al coeficiente de correlación policórica:

$$\omega_o = \frac{(\sum \lambda_j)^2}{k + 2\sum \rho^*_{j<j'}} \qquad (6)$$

Para medidas esencialmente tau-equivalentes, puede utilizarse el coeficiente alfa ordinal (Elosua y Zumbo, 2008; Gadermann et al., 2012; Zumbo, Gadermann, y Zeisser, 2007), simplificando el numerador en coherencia con la Ecuación 4:

$$\omega_o = \alpha_o = \frac{k^2\lambda^2}{k + 2\sum \rho^*_{j<j'}} \qquad (7)$$

Como ocurre con α para los datos cuantitativos, el coeficiente alfa ordinal sólo se aconseja si el modelo subyacente a los datos es el de medidas esencialmente tau-equivalentes, mientras que el coeficiente omega ordinal se aconseja si el modelo subyacente es el de medidas congenéricas (Gadermann et al., 2012; Napolitano, Callina, y Mueller, 2013). También el sesgo de alfa ordinal sería más grave cuando en el modelo no se especificaran los errores correlacionados, ya que se incluirán en el numerador como parte de la varianza verdadera tal como sucedía con los datos cuantitativos.

Estos coeficientes ordinales tienen ventajas porque constituyen una generalización directa de los coeficientes omega lineales, pero también presentan limitaciones, ya que no evalúan la fiabilidad de la suma o promedio de los ítems observados (Y_j), sino la de las respuestas latentes continuas subyacentes (Y_j^*). Si los investigadores están interesados en la fiabilidad de la suma de los ítems, una opción mejor es calcular la fiabilidad no lineal basada en SEM (Green y Yang, 2009; Yang y Green, 2015).Conceptualmente la ecuación es cercana a la de omega ordinal, pero la probabilidad acumulada de los umbrales bajo la ley normal se incluye tanto en el numerador como en el denominador para re-expresar las varianzas de la puntuación verdadera y del error en la métrica de la suma de los ítems observados. El cálculo es complejo, por lo que los autores proporcionan el código para el *software* SAS en un apéndice (Green y Yang, 2009). Utilizando R también se puede obtener el IC para el coeficiente de fiabilidad no lineal basado en SEM (Kelley y Pornprasertmanit, 2016) siempre que el modelo de medidas congenéricas sea aceptable.

En el caso de que los investigadores decidieran ajustar un modelo de medida basándose en la TRI, la fiabilidad de consistencia interna también podría derivarse de los parámetros estimados. Como en el caso del AFC, los modelos de TRI proporcionan estimaciones de los parámetros del ítem y de la distribución de la variable latente que permiten cuantificar las varianzas de las puntuaciones totales, de las puntuaciones verdaderas y de los errores tal como se conciben en la TCT (Kim y Feldt, 2010 y sus referencias), y a partir de estas varianzas, se puede estimar la fiabilidad como la proporción de la varianza de la puntuación verdadera contenida en la varianza total, tal y como como se define en la Ecuación 1.

Para datos dicotómicos unidimensionales y un modelo de medidas congenéricas, Dimitrov (2003) desarrolló la estimación puntual de la fiabilidad de consistencia interna para modelos de uno, dos o tres parámetros, siempre que los ítems hubieran sido previamente calibrados. Dimitrov propuso utilizar cálculos aproximados para evitar complejidades computacionales y así facilitar que las fórmulas se pudieran implementar en una hoja de cálculo, en programas estadísticos básicos o programar en R. Posteriormente, Raykov, Dimitrov y Asparouhov (2010) desarrollaron estas ideas incorporándolas a un método que permite simultáneamente calibrar los ítems y calcular el IC del coeficiente de fiabilidad de consistencia interna para la suma de ítems tanto para modelos de uno como de dos parámetros. Como de costumbre, los autores proporcionan la sintaxis en Mplus para estimar los parámetros del modelo y el IC de la fiabilidad en una sola ejecución.

Aplicación a cuatro escenarios prácticos

Para ilustrar los conceptos anteriores, presentamos cuatro ejemplos que simulan situaciones de investigación aplicada en las que

se pretende representar un único constructo mediante la suma o promedio de respuestas a un conjunto de ítems. En cada uno de los cuatro escenarios, analizamos las respuestas simuladas de 600 personas a 6 ítems en una escala Likert de cinco puntos. La fuente de conocimiento previo acerca de los modelos que son compatibles con las respuestas analizadas constituye una diferencia importante con respecto a la investigación aplicada. En un contexto aplicado el necesario conocimiento previo procede de la teoría subyacente y de los estudios anteriores, mientras que en nuestros ejemplos dicho conocimiento procede de los modelos utilizados para simular los datos.

En el Caso 1, el modelo de medición subyacente a los datos es el de medidas esencialmente tau-equivalente con cargas factoriales altas cercanas a .65, y distribuciones de respuesta simétricas. Se espera, en consecuencia, que la exploración de los datos sugiera seguir con un análisis cuantitativo, que el modelo de medida de datos esencialmente tau-equivalente sea el que mejor ajuste, y que el valor del coeficiente omega sea igual al valor de α. En el Caso 2, el modelo subyacente es el de medidas congenéricas con distribuciones de respuesta simétricas y cargas factoriales homogéneamente altas. En consecuencia, se espera que las respuestas a los ítems puedan ser tratadas como cuantitativas, que el modelo con mejor ajuste sea el de medidas congenéricas y que el valor de α esté próximo al de omega debido a las cargas factoriales altas y con valores homogéneos. El modelo subyacente al Caso 3 tiene distribuciones de respuesta simétricas, cargas factoriales muy variables y tres ítems con errores correlacionados. Así pues, se espera que la descripción de los datos apoye un análisis cuantitativo posterior, que el mejor modelo sea el de medidas con errores correlacionados y que el valor de α sea mayor que el de omega, sobrestimando la fiabilidad debido principalmente al hecho de que omega corrige la correlación entre los errores mientas que α trata esa correlación como varianza verdadera. Por último, en el Caso 4, el modelo subyacente es el de medidas congenéricas, con fuertes efectos de techo en las distribuciones de respuesta y con cargas factoriales muy variables. En este caso, se espera que el análisis descriptivo sugiera tratar los datos como ordinales y que el modelo de medidas congenéricas muestre el mejor ajuste. En cuanto a los dos coeficientes de fiabilidad, se espera que muestren una diferencia considerable puesto que el coeficiente alfa ordinal estima la fiabilidad de las respuestas latentes esencialmente tau-equivalentes, mientras que el coeficiente de fiabilidad no lineal basado en SEM estima la fiabilidad de las respuestas congenéricas observadas.

Todos los análisis se realizaron con R. La Fase 1, de descripción de los datos, con los paquetes *reshape2* (Wickham, 2007) y *psych* (Revelle, 2016) para calcular los porcentajes de respuesta y otros estadísticos descriptivos, así como los coeficientes de correlación de Pearson o policórica cuando era apropiado. En la Fase 2 se utilizó la función *cfa* del paquete *lavaan* (Rosseel, 2012) para analizar los modelos de medida anidados. En los tres primeros casos, con datos cuantitativos, se eligió la estimación ML, mientras que en el Caso 4 se escogió la estimación WLSMV adecuada para datos ordinales. Para facilitar la comparación de los resultados proporcionados por ambos coeficientes, en la Fase 3 se obtuvieron los coeficientes alfa y omega para los modelos de medida más parsimoniosos y con mejor ajuste usando la función *reliability* del paquete *semTools* (semTools Contributors, 2016). Cuando el paquete lo permitía, se calcularon los intervalos de confianza del 95% utilizando la función *ci.reliability* del paquete *MBESS* (Kelley y Pornprasertmanit, 2016). La toma de decisiones se basó en los criterios descritos en los apartados anteriores. Las bases de datos de los ejemplos están disponibles en http://ddd.uab.cat/record/173917 y la sintaxis utilizada para su resolución se encuentra en el apéndice A y en el apéndice B de este artículo.

La Tabla 1 presenta los estadísticos descriptivos univariados y bivariados para todos los escenarios. En el Caso 1, las categorías centrales mostraron el mayor porcentaje de respuestas y se observaron efectos de techo o suelo. Los valores de asimetría se situaron entre -0.11 y 0.10 y los de curtosis entre -0.29 y -0.64, de modo que, aunque procedentes de las respuestas a una escala de Likert de cinco puntos, los datos fueron tratados como cuantitativos. Todos los coeficientes de correlación de Pearson fueron positivos, homogéneos y se situaron entre .31 y .47. Por lo tanto, decidimos utilizar el estimador ML para probar los dos modelos de medida plausibles; el de medidas congenéricas *versus* el de medidas esencialmente tau-equivalentes. Los resultados se presentan en las dos primeras líneas de la Tabla 2. El modelo más restringido que se probó, el modelo de medidas esencialmente tau-equivalentes, mostró un buen ajuste a los datos, χ^2 (14) = 22.02, p = .078, CFI = .992 , TLI = .991, RMSEA = .031. Como la diferencia de χ^2 con el modelo de medidas congenéricas, más flexible, no fue estadísticamente significativa, χ^2 (5) = .09, p = .999, elegimos el de medidas esencialmente tau-equivalentes en aplicación del principio de parsimonia. Por lo tanto, se cumplieron todos los supuestos para que α (véase la Ecuación 4) fuera un buen estimador de la fiabilidad de la consistencia interna. Como era de esperar, la estimación de α = .809 es la misma que la estimación de omega. La consistencia interna de la suma o promedio de los ítems en el Caso 1 se encuentra dentro de valores aceptados habitualmente, con valores del 95%IC entre .784 y .831.

La exploración de los datos en el Caso 2 también nos llevó a tratarlos como cuantitativos. De hecho, los estadísticos descriptivos de la Tabla 1 muestran las frecuencias en una escala de cinco puntos sin efectos de techo o suelo, con valores de asimetría y curtosis no superiores, en valor absoluto a 0.19 y 0.85 respectivamente y coeficientes de correlación entre ítems homogéneos, con un rango entre .26 y .53. En consecuencia, se probaron los modelos de medidas congenéricas y esencialmente tau-equivalentes usando el estimador ML. Como se observa en la Tabla 2, cuando se impuso la restricción de cargas factoriales iguales se obtuvo un ajuste inaceptable (medidas esencialmente tau-equivalentes, con χ^2 (14) = 46.78, p <.001, CFI = .969, TLI = .967, RMSEA = .062. Se observó una considerable mejoría en el ajuste cuando se permitió, con el modelo más flexible de medidas congenéricas, que las cargas factoriales fueran diferentes entre los ítems, con χ^2 (9) = 20.46, p = .015, CFI = .989, TLI = .982 y RMSEA =.046. Por otra parte, la diferencia χ^2 entre ambos modelos fue estadísticamente significativa, χ^2 (5) = 26.32, p < .001, lo que indica un mejor ajuste del modelo de medidas congenéricas. Por lo tanto, en este caso, las estimaciones de consistencia interna se deben obtener utilizando el coeficiente omega (véase la Ecuación 2). Sin embargo, como ya se anticipó, tanto el coeficiente omega (.823) como α (.820) mostraron valores similares ya que todas las cargas factoriales eran altas y ho-

mogéneas (entre .60 y .83). Los valores mínimos de los 95%IC de ambos coeficientes fueron claramente superiores a los estándares habituales, lo que constituye evidencia a favor de la consistencia interna de las puntuaciones.

De nuevo, en el Caso 3, todos los estadísticos descriptivos sugirieron analizar los datos como cuantitativos. Las distribuciones de respuesta en cinco categorías no mostraron respuestas extremas y los índices de asimetría y curtosis no fueron mayores, en valor absoluto, de 0.17 y 0.51 respectivamente (ver Tabla 1) y por lo tanto el estimador ML fue considerado apropiado. Sin embargo, como se esperaba, los coeficientes de correlación entre ítems no fueron homogéneos, ya que se observaron correlaciones muy altas, superiores a .78, entre tres de los ítems (Y4, Y5, Y6), mientras que las correlaciones restantes tomaron valores entre bajos y moderados, de .05 a .43. La agrupación especial entre estos tres ítems fue modelizada mediante errores correlacionados entre ellos.

Tabla 1. Resultados de la Fase 1 en cuatro escenarios prácticos: Estadísticos descriptivos univariados y coeficientes de correlación.

		Estadísticos univariados									Correlaciones				
		%1	%2	%3	%4	%5	*M*	*DE*	*s*	*k*	Y1	Y2	Y3	Y4	Y5
Caso1															
	Y1	14.67	26.17	38.50	16.17	4.50	2.70	1.05	0.10	-0.51					
	Y2	9.17	17.67	37.00	22.33	13.83	3.14	1.14	-0.09	-0.64	.31				
	Y3	10.83	25.00	42.33	18.50	3.33	2.79	0.98	-0.04	-0.38	.43	.43			
	Y4	3.50	18.33	35.83	30.50	11.83	3.29	1.01	-0.11	-0.54	.47	.33	.42		
	Y5	2.83	14.67	43.83	26.50	12.17	3.31	0.96	0.00	-0.29	.41	.42	.46	.43	
	Y6	4.67	20.33	39.17	28.17	7.67	3.14	0.98	-0.09	-0.41	.42	.40	.43	.46	.46
Caso2															
	Y1	17.00	27.17	32.67	17.50	5.67	2.68	1.12	0.17	-0.70					
	Y2	7.33	19.33	36.17	24.17	13.00	3.16	1.11	-0.07	-0.63	.26				
	Y3	14.50	23.67	35.33	19.17	7.33	2.81	1.13	0.07	-0.68	.39	.37			
	Y4	7.50	19.33	28.17	27.67	17.33	3.28	1.18	-0.19	-0.85	.47	.33	.46		
	Y5	5.50	18.00	35.67	24.33	16.50	3.28	1.11	-0.09	-0.68	.42	.42	.46	.50	
	Y6	7.33	21.67	32.17	28.00	10.83	3.13	1.10	-0.11	-0.70	.42	.44	.48	.53	.52
Caso3															
	Y1	14.67	26.17	38.50	16.17	4.50	2.70	1.05	0.10	-0.51					
	Y2	6.17	19.33	39.50	24.50	10.50	3.14	1.04	-0.05	-0.46	.05				
	Y3	3.50	16.67	40.50	25.67	13.67	3.29	1.01	-0.02	-0.47	.28	.25			
	Y4	4.67	16.67	36.67	30.00	12.00	3.28	1.03	-0.17	-0.46	.25	.19	.36		
	Y5	11.00	25.33	39.33	18.50	5.83	2.83	1.04	0.07	-0.45	.17	.20	.29	.79	
	Y6	6.83	18.17	40.83	25.50	8.67	3.11	1.02	-0.12	-0.35	.27	.26	.43	.86	.83
Caso4															
	Y1	2.17	5.33	9.83	21.33	61.33	4.34	1.00	-1.56	1.72					
	Y2	2.00	5.17	11.50	20.00	61.33	4.34	1.00	-1.49	1.46	.19				
	Y3	0.83	4.33	10.00	20.83	64.00	4.43	0.90	-1.58	1.85	.33	.41			
	Y4	0.67	3.50	12.83	17.67	65.33	4.43	0.89	-1.49	1.41	.39	.47	.64		
	Y5	1.50	3.83	12.00	21.83	60.83	4.37	0.94	-1.50	1.67	.39	.44	.57	.61	
	Y6	1.67	4.17	12.33	20.17	61.67	4.36	0.96	-1.50	1.58	.44	.46	.39	.54	.44

Nota. %1 a %5: porcentajes de respuesta a cada categoría; *s* = asimetría; *k* = curtosis. Coeficientes de correlaciones de Pearson (Caso1, Caso2 y Caso3) o policóricas (Caso4).

Como se muestra en la Tabla 2, el ajuste del modelo de medidas esencialmente tau-equivalentes no fue aceptable, χ^2 (14) = 608.25, p <.001, CFI = .669, TLI = .645, RMSEA = .266). Los índices de bondad de ajuste para el modelo de medidas congenéricas, aunque mejor, χ^2 (9) = 78.19, p < .001, CFI = .961, TLI = .936, RMSEA = .113, tampoco fueron aceptables, con la excepción del CFI. Modelizando las altas correlaciones entre los ítems Y4, Y5 e Y6 como correlaciones entre sus errores, se observaron buenos índices de ajuste, χ^2 (6) = 13.82, p = .032, CFI = .996, TLI = .989, RMSEA = .047, a excepción del valor de χ^2 estadísticamente significativa. Además, se encontró una diferencia estadísticamente significativa con el modelo de medidas congenéricas, χ^2 (3) = 64.37, p < .001, lo que indica que el modelo con errores correlacionados presenta un ajuste significativamente mejor que dicho modelo.

En coherencia con este modelo de medida, se obtuvo la estimación de la consistencia interna con el coeficiente omega corrigiendo por la presencia de errores correlacionados (ver Ecuación 5). El valor observado de .560, por debajo de los estándares habituales, conlleva la conclusión de que la suma de estos ítems no es fiable, una conclusión consistente con el resultado de la Fase 2 donde la unidimensionalidad de los datos observados quedó seriamente cuestionada, llevando ambos datos a la conclusión de que el uso de la puntuación por suma de los ítems no es apropiado en el Caso 3. El hecho de que el modelo de medidas esencialmente tau-equivalentes no se ajuste a los datos, y especialmente la presencia de ítems con errores correlacionados, debe desalentar el uso de α para la estimación de la fiabilidad de consistencia interna. A pesar de ello, en la Tabla 2 se incluye el valor de α con el fin de ilustrar el cambio radical que se produciría en la conclusión en caso de que se utilizara este coeficiente, ya que su valor de .773 llevaría fácilmente a la creencia incorrecta de que los ítems son consistentes.

En el Caso 4 se observan efectos techo muy claros, todos los ítems acumulan más del 60% de las respuestas en la última categoría tal y como se observa en la Tabla 1. A pesar de que los valores de asimetría y curtosis no fueron particularmente destacados, sí tomaron valores fuera del rango entre -1 y 1. En consecuencia, dado que los datos provienen de una escala de respuesta tipo Likert, las distribuciones de las respuestas sugieren la conveniencia de considerarlos como ordinales. Por esta razón, se obtuvieron los coeficientes de correlación policórica, que se situaron en un amplio rango de valores entre .19 y .64, sin que se observara ninguna agrupación particular entre los ítems. Por la misma razón, para ajustar a los datos los modelos de medidas congénericas y esencialmente tau-equivalentes, se utilizó el estimador WLSMV.

Tabla 2. Resultados de las Fase 2 y Fase 3 en cuatro escenarios prácticos: Resultados principales de los modelos de medida y coeficientes de fiabilidad.

	Fase 2							Fase 3		
Caso (Modelo simulado)	Modelo ajustado	Pesos factoriales	χ^2	df	*p*	CFI	TLI	RMSEA [IC95%]	Alfa [IC95%]	Omega/ Fiabilidad no lineal [IC95%]
Caso 1 (TM)										
	TM	.66	22.02	14	.078	.992	.991	.031 [.000, .054]	.809 [.784, .831]	.809 [.786, .830]
	CM	.65, .65, .66, .66, .66, .66	21.93	9	.009	.987	.978	.049 [.023, .075]		
Caso 2 (CM)										
	TM	.75	46.78	14	<.001	.969	.967	.062 [.043, .083]		
	CM	.60, .66, .74, .79, .82, .83	20.46	9	.015	.989	.982	.046 [.019, .073]	.820 [.797, .842]	.823 [.799, .845]
Caso 3 (CE)										
	TM	.85	608.25	14	<.001	.669	.645	.266 [.248, .284]		
	CM	.26, .28, .43, .90, .92, .99	78.19	9	<.001	.961	.936	.113 [.091, .137]		
	CE	.36, .41, .47, .57, .67, .68	13.82	6	.032	.996	.989	.047 [.013, .079]	*.773*	.560 [,]
Caso4 (CM)										
	TM	0.69	110.11	14	<.001	.950	.946	.102 [.084, .121]		
	CM	.49, .58, .66, .74, .74, .85	35.15	9	<.001	.994	.989	.045 [.018, .072]	*.830*	.777 [.738, .809]

Nota. Todas las cargas factoriales están estandarizadas. Cursiva: coeficiente alfa estimado de forma incorrecta e incluido con propósito comparativo. TM = Medidas esencialmente tau-equivalentes; CM = medidas congenéricas CE = medidas con errores correlacionados; CFI = índice de ajuste comparativo; TLI = índice de Tucker-Lewis; RMSEA = error cuadrático medio de aproximación; IC = intervalo de confianza. [,] = IC no disponible.

Como se observa en la Tabla 2, el ajuste al modelo de medidas esencialmente tau-equivalentes fue inaceptable, χ^2 (14) = 100.78, $p < .001$, CFI = .950, TLI = .946, RMSEA = .102), mientras que fue bueno para el modelo de medidas congenéricas, χ^2 (9) = 20.07, $p = .018$, CFI = .994, TLI = .989, RMSEA = .045, con la excepción del valor de χ^2 estadísticamente significativo. Por lo tanto, elegimos el modelo de medidas congenéricas como el más adecuado para estos datos. En coherencia con el modelo de medida ajustado, el estimador de la fiabilidad de consistencia interna adecuado fue el coeficiente de fiabilidad no lineal basado en SEM (véase Green y Yang, 2009) con un valor de .777, 95%IC [.739, .808]. Todos estos valores están dentro de los estándares aceptados en el proceso de desarrollo de una escala. Con el coeficiente alfa ordinal (véase la Ecuación 7) se habría obtenido un valor de .830, claramente superior aunque incorrecto ya que no se cumple la condición de tau-equivalencia. Por otra parte, el coeficiente alfa ordinal estimaría la fiabilidad de la suma de las variables de respuesta latentes y no la fiabilidad de la suma de las respuestas observadas.

Generalización a modelos de medida y diseños complejos

En este apartado se generalizarán tanto la lógica como los resultados de las secciones anteriores a los casos de medidas esencialmente unidimensionales, escalas multidimensionales, a diseños multinivel y a datos con valores faltantes, así como al uso de los coeficientes de fiabilidad en el desarrollo y revisión de escalas.

Fiabilidad de medidas esencialmente unidimensionales

Todos los modelos discutidos hasta ahora comparten el supuesto de que los ítems miden un solo constructo. La presencia de correlación entre los ítems después de controlar el factor común, como en el Caso 3, se ha tratado como una anomalía a corregir. Sin embargo, este es un caso particular de un tema más general. Cada ítem puede medir tanto el constructo deseado

como otros factores que el investigador considera espurios. Los motivos para la presencia de dichos factores son variados e incluyen características del cuestionario, como la formulación directa o inversa de los ítems o la presencia de *testlets*, y también sesgos de respuesta tales como la deseabilidad social, el afecto negativo o la aquiescencia (e.g., Conway y Lance, 2010; Lance, Dawson, Birkelbach, y Hoffman, 2010; Spector, 2006). En este apartado utilizaremos el concepto de la unidimensionalidad esencial acuñado por Stout (1987; véase también Raykov y Pohl, 2013) para tratar de una manera más general los cuestionarios que miden predominantemente un factor, pero donde pueden identificarse factores espurios adicionales formados por subgrupos de ítems.

Cuando se sospecha que hay fuentes espurias de variabilidad, durante la Fase 1 del análisis el analista podrá observar cuidadosamente la matriz de correlaciones en busca de agrupamientos de ítems, tal y como hemos visto en el Caso 3. El análisis formal, sin embargo, se llevará a cabo en la Fase 2. La especificación de un modelo de medida del tipo bifactor (e.g., Reise, 2012) es particularmente útil para determinar la unidimensionalidad esencial. En este modelo, representado en la Figura 4, se permite que cada ítem pese en un factor general y también en un factor específico que podría ser espurio. Se puede definir más de un factor específico para adecuarse a diversos grupos de ítems. Si el modelo bifactor ajusta a los datos y los investigadores creen que los factores específicos son espurios, entonces deben incluir este conocimiento en la Fase 3 de estimación de la fiabilidad. El coeficiente apropiado, denominado omega jerárquico por Zinbarg et al. (2005) y aplicado al diagrama de la Figura 4 es:

$$\omega_h = \frac{(\sum \lambda_{gj})^2}{(\sum \lambda_{gj})^2 + (\sum \lambda_{sj})^2 + \sum \sigma^2_{\varepsilon_j}} \qquad (8)$$

En el numerador, la varianza verdadera se deriva del factor general, mientras que la varianza debida a factores específicos se trata como la varianza del error al integrarse únicamente en el denominador. Esta formulación excluye del numerador del coeficiente de fiabilidad toda la varianza espuria, ya sea atribuible a factores de método, a especificidades de los ítems, al proceso de respuesta o a variación aleatoria. Siempre que el modelo se ajuste a los datos, en el denominador se puede utilizar la suma de las varianzas y covarianzas observadas, tal como discutimos al presentar la Ecuación 3. Este coeficiente omega jerárquico constituye una especificación más general de la Ecuación 5 tal como se explica en Gu et al. (2013) para datos cuantitativos y en Yang y Green (2011) para datos ordinales.

Si se dispone de conocimiento previo sobre las posibles fuentes de varianza espurias, se puede especificar y estimar el modelo bifactor de forma confirmatoria utilizando el paquete *lavaan* de R o el software comercial Mplus. Si se quiere aportar pruebas de unidimensionalidad en ausencia de conocimientos previos sobre fuentes particulares de varianza espuria, se puede estimar un modelo bifactor exploratorio utilizando las rotaciones de Schmid-Leiman o de Jennrich-Bentler (e.g., Mansolf y Reise, 2016) mediante el paquete *psych* de R o el *software* comercial Mplus. Esta aproximación exploratoria de tipo general es más adecuada que la práctica habitual de la re-especificación de parámetros basada en los índices de modificación derivados de un modelo de medidas congenéricas con mal ajuste (e.g., Brown, 2015; Hoyle, 2012).

Por muy razonable que parezca, ésta es sólo una de las dos visiones de la fiabilidad de consistencia interna basadas en SEM (Zinbarg et al., 2005). Ambas se derivan del hecho que conceptualmente, en el análisis factorial, la puntuación observada puede descomponerse en cuatro partes, a saber, un factor general en el que pesarían todos los ítems, factores formados por grupos de ítems (factores de grupo), factores específicos de cada ítem y variación aleatoria. En cambio, en la TCT, la puntuación observada se descompone sólo en dos partes: una verdadera y una que contiene error. Existe consenso en que el factor general forma parte de la varianza verdadera y la variación aleatoria forma parte de la varianza del error. Los factores de grupo debidos a los diferentes contenidos también se consideran parte de la varianza verdadera y, si están presentes, convertirán el cuestionario en multidimensional. Las diferentes conceptualizaciones de la fiabilidad proceden de considerar los factores de grupo espurios y los factores específicos bien como parte de la varianza verdadera o bien de la varianza de error. La respuesta que se da al calcular omega jerárquico es que la variabilidad espuria y específica, que no forman parte del constructo, forman parte de la varianza de error.

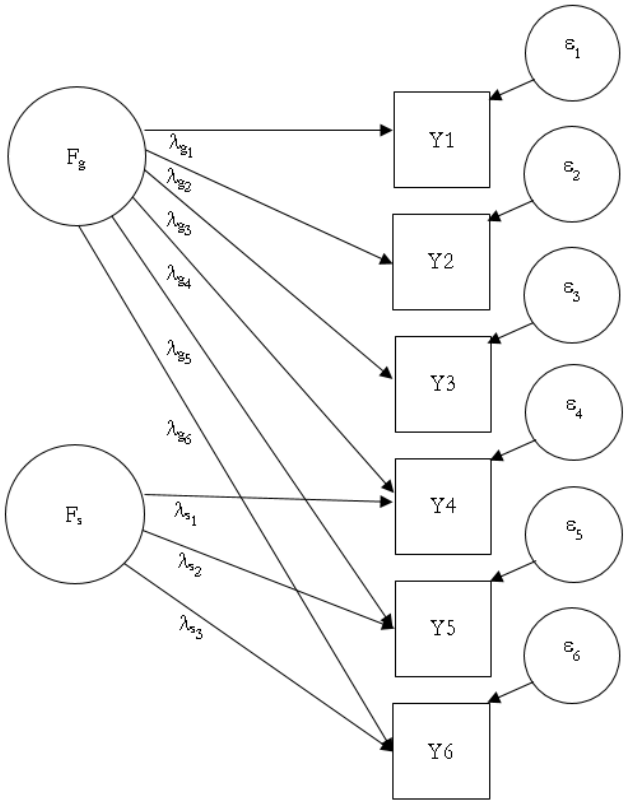

Figura 4. Modelo bifactor con tres ítems que muestran un efecto de método.

Por otra parte, una conceptualización más clásica de la fiabilidad sostiene que todos los factores sistemáticos contribuyen a las correlaciones entre los ítems y, lo que es más importante, a las correlaciones con las variables externas; únicamente los errores aleatorios tienen el efecto de atenuar estas correlaciones. En

consecuencia con este razonamiento, la fiabilidad de la suma o promedio de ítems debería incluir toda la variación sistemática, ya sea debida a factores de contenido, de método o específicos. Y con más razón si es que se quiere seguir considerando el coeficiente de fiabilidad como límite superior del coeficiente de validez predictiva. Los investigadores que se identifican con esta posición tomarán partido por calcular la fiabilidad de consistencia interna a través del coeficiente que Zinbarg et al. (2005) denomina omega total y cuya expresión aplicada al ejemplo de la Figura 4 es:

$$\omega_t = \frac{(\sum \lambda_{gj})^2 + (\sum \lambda_{sj})^2}{(\sum \lambda_{gj})^2 + (\sum \lambda_{sj})^2 + \sum \sigma_{\varepsilon_j}^2} \qquad (9)$$

En omega total el factor de grupo se considera parte de la varianza verdadera y por lo tanto se incluye en el numerador. Como advirtió Bentler (2009), la decisión de considerar factores espurios y específicos como parte de la varianza verdadera o de error debe depende de los objetivos de los investigadores. En una sola administración de un cuestionario destinado a medir un constructo, esta discusión está circunscrita a los posibles factores de grupo espurios, ya que los factores específicos para cada ítem no pueden distinguirse de la variación aleatoria.

En este contexto, la propuesta de Green y Yang (2015) de publicar tanto omega jerárquica como omega total parece razonable. Esto permite no sólo evaluar la fiabilidad bajo las dos conceptualizaciones, sino que también proporciona una forma sencilla de evaluar la unidimensionalidad. Un alto parecido entre ambos valores proporcionaría evidencia favorable para la unidimensionalidad, ya que los factores espurios no generarían mucha variación sistemática. Ambos valores pueden obtenerse derivándolos de los parámetros estimados mediante un modelo bifactor confirmatorio utilizando Mplus o los paquetes *lavaan* y *semTools* de R. Sus versiones exploratorias pueden obtenerse utilizando la función *omega* del paquete *psych* de R.

Adicionalmente, en diseños longitudinales se pueden identificar las especificidades de los ítems para cuyo tratamiento se han propuesto varias soluciones. McCrae (2014) mantuvo la postura de que sería más apropiado atender a la fiabilidad test-retest ya que incluye sin duda alguna la especificidad de los ítems como varianza verdadera, mientras que Bentler (2016) propuso índices de consistencia interna mejorados atendiendo a la especificidad, y Raykov y Marcoulides (2016b) proporcionaron la justificación y la sintaxis para la estimación de la varianza específica mediante SEM con el paquete Mplus.

Finalmente, queremos resaltar que las formulaciones de la Ecuación 2 hasta la Ecuación 9 son pertinentes para estimar la fiabilidad cuando se planea incorporar la suma o el promedio de los ítems en posteriores análisis de tipo predictivo, de comparación de grupos o de tipo longitudinal. Ambas son combinaciones lineales obtenidas atribuyendo pesos iguales a todos los ítems. Sin embargo, si el objetivo es estudiar las relaciones entre variables latentes en un modelo SEM, los constructos se medirían a través de la combinación lineal óptima de sus indicadores, de manera que su fiabilidad de consistencia interna se estimaría de forma más adecuada con el coeficiente H (Hancock y Mueller, 2001, 2013) también conocido como de fiabilidad máxima (Raykov, 2012):

$$H = \frac{\sum[\lambda_j^2/(1-\lambda_j^2)]}{1+\sum[\lambda_j^2/(1-\lambda_j^2)]} \qquad (10)$$

La relación entre la comunalidad (λ_j^2)y la especificidad ($1-\lambda_j^2$) de cada ítem es el núcleo del coeficiente H. El coeficiente puede interpretarse como la proporción máxima de varianza del constructo teórico que puede ser explicada por sus indicadores, o en otras palabras, como la fiabilidad de la combinación lineal óptima entre ítems. Entre sus propiedades, destacamos que el valor de H es igual o mayor que la fiabilidad del ítem más fiable, no depende del signo de las cargas factoriales ni disminuye cuando aumenta el número de ítems. La Ecuación 10 sólo es adecuada si el modelo de medidas esencialmente tau-equivalentes o el de medidas congenéricas se ajustan a los datos. Si se utiliza un modelo de medida con errores correlacionados, este coeficiente debería corregirse consecuentemente (Gabler y Raykov, 2016). Por otra parte, si la puntuación latente se calculara partiendo de un modelo TRI, su fiabilidad debería estimarse en consecuencia (e.g., Cheng, Yang y Liu, 2012).

Sin embargo, la estimación de efectos estructurales entre las variables latentes a partir de la metodología SEM tiene sus propios inconvenientes, ya que el uso de la combinación lineal óptima proporciona medidas que dependen de la muestra y del particular momento temporal en que se han obtenido (Raykov, Gabler y Dimitrov, 2016). Estos autores sugieren usar esta medida más compleja sólo si es absolutamente necesario, es decir, cuando la fiabilidad de la combinación lineal óptima (coeficiente H, Ecuación 10) es estadísticamente mayor que la fiabilidad de la combinación lineal con todos los ítems igualmente ponderados (coeficiente omega, Ecuación 2).

Fiabilidad en medidas multidimensionales

Hasta ahora, nos hemos centrado en escalas de medida unidimensionales, quizás afectadas por factores espurios, dejando fuera una amplia gama de otros modelos de medida muy útiles. Por ejemplo, ¿cómo calcular la fiabilidad de consistencia interna de las puntuaciones derivadas de múltiples factores quizás correlacionados? Este es el caso de numerosas escalas en el ámbito de las ciencias sociales. Un ejemplo de esto sería una medida de motivación, que incluirá al menos una escala de motivación intrínseca, otra de motivación orientada externamente y tal vez, una tercera de falta de motivación. Por razones teóricas, se espera que estos constructos estén correlacionados entre sí, algunos positivamente y otros negativamente. Y también, ¿cómo calcular la fiabilidad de consistencia interna de las puntuaciones derivadas de un modelo de medida jerárquico, con un factor general y algunos factores específicos con contenido interpretable? Un ejemplo clásico es el modelo de inteligencia con un factor general y otros específicos de inteligencia verbal, lógica, manipulación, etc. O aún más difícil, ¿qué hacer si toda la escala está compuesta por ítems complejos? Nos referimos a ítems que muestran sistemáticamente cargas cruzadas menores en varios factores además de una carga factorial elevada en el factor deseado (Marsh et al., 2010), como ocurre, por ejemplo, en pruebas de personalidad tales como el test *Big Five*.

Para determinar la fiabilidad en estas estructuras sigue siendo útil seguir el procedimiento de las tres fases analíticas descri-

to anteriormente. Probablemente, durante la Fase 1, exploratoria, se puedan observar algunos grupos de variables, pero la prueba formal para evaluar la multidimensionalidad se llevará a cabo en la Fase 2, al estudiar el ajuste del modelo de medida a los datos. La evidencia empírica puede favorecer un modelo con múltiples factores ortogonales o con múltiples factores correlacionados, un modelo bifactor, o incluso un modelo factorial de segundo orden (e.g., Ntoumanis, Mouratidis, Ng, y Viladrich, 2015). Al igual que en los casos unidimensionales, es necesario que el modelo de medida adoptado tenga sentido teórico y se ajuste a los datos. Las reglas para resolver la Fase 3, el cálculo del coeficiente omega aplicado a la escala multidimensional que se está analizando, se encontrarán fácilmente o podrán derivarse de la literatura actual. Para dar algunos ejemplos, Black, Yang, Beitra y McCaffrey (2015) explican cómo calcular la fiabilidad en modelos factoriales de segundo orden y bifactor aplicados a una prueba de inteligencia; Gignac (2014) la fiabilidad de un factor general procedente de una escala multidimensional; Green y Yang (2015) la fiabilidad de factores específicos, aplicable al estudio de la fiabilidad de modelos con factores correlacionados; Raykov y Marcoulides (2012) la fiabilidad de escalas multidimensionales y su validez relacionada con un criterio; Cho (2016) el cálculo del coeficiente omega en varios modelos multidimensionales para datos cuantitativos; o Rodríguez, Reise, y Haviland (2016) el cálculo e interpretación de coeficientes de fiabilidad derivados de modelos bifactor.

Fiabilidad en diseños complejos: datos faltantes y diseños multinivel

Otra cuestión que debe abordarse es el modo de tratar las características de los datos derivadas del diseño de la investigación y también del estudio de campo. A este respecto, los investigadores pueden preguntarse: ¿Cómo calcular la consistencia interna cuando las personas participantes están anidadas en estructuras como aulas, escuelas, equipos o empresas, proporcionando así datos multinivel? ¿Y cuando los datos están incompletos? Nuestra respuesta sería que el procedimiento analítico en tres fases sigue funcionando bajo estas condiciones. Es decir, siempre que se estimen correctamente los parámetros del modelo de medida, una consecuencia natural será que la estimación de la fiabilidad basada en estos parámetros será correcta.

En el caso de los datos multinivel, se puede añadir en la Fase 1 del análisis el coeficiente de correlación intraclase con el fin de evaluar la magnitud del efecto de agrupamiento. En la Fase 2, se deberá usar una corrección apropiada para dicho efecto, por ejemplo, agregando en Mplus la siguiente línea de sintaxis: *type = complex*. Una vez que se hayan estimado los parámetros correctamente, en la Fase 3, se puede obtener el coeficiente de fiabilidad utilizando las ecuaciones presentadas en los apartados anteriores. En el documento de Raykov et al. (2015) se puede encontrar una aplicación a los datos multinivel. Estos autores presentan todos los detalles necesarios, incluyendo la sintaxis en Mplus, para calcular α con la estimación MLR y los errores estándar corregidos por el efecto de agrupamiento. En Raykov y Marcoulides (2014) puede encontrarse una generalización útil para el análisis de poblaciones heterogéneas.

En cambio, enfrentarse a datos incompletos requiere estrategias más matizadas. En primer lugar, se debe tener extrema precaución en el diseño de obtención de datos y durante el estudio de campo, porque la mejor manera de tratar los datos faltantes es que, al llegar al estadio de análisis, no los haya. Con todo, si se debe analizar una base de datos incompleta se necesitarán métodos específicos. Los detalles concretos superan los objetivos de este documento, aunque se desarrollarán de forma esquemática y se pueden encontrar ampliados en la literatura metodológica (e.g., Enders, 2010, 2013; Graham, 2009). En la Fase 1, se deberá evaluar la proporción de datos faltantes, ya que en pequeñas cantidades no tienen consecuencias graves en los análisis posteriores. Si está presente una proporción moderada, se recomienda explorar y analizar su estructura, ya que si son al azar tampoco tienen efectos importantes sobre los resultados del SEM si se utiliza un estimador de parámetros ML. Finalmente, las estrategias más elaboradas serían necesarias en caso de que la proporción de datos faltantes sea grande y/o no al azar. Un ejemplo de datos faltantes no al azar se puede encontrar en la evaluación de la eficacia de un tratamiento, cuando algunos participantes abandonan el tratamiento debido a que no ha cumplido sus expectativas. Una de estas estrategias, la inclusión de variables auxiliares, se explica en el artículo de Raykov y Marcoulides (2016a) donde los autores incluyen la sintaxis de Mplus para calcular la fiabilidad y el IC de las puntuaciones de una escala en estas circunstancias. Otra opción la ofrece el paquete *coefficientalpha* desarrollado en R y documentado en Zhang y Yuan (2016) que permite estimar los coeficientes alfa y omega y su IC en presencia de datos faltantes y de casos atípicos de forma coherente con la metodología de análisis expuesta en nuestro trabajo aunque se aplica sólo a un reducido rango de modelos de medida.

Cambio en la fiabilidad por revisión de una escala

Otro uso muy generalizado de α ha sido el tratarlo como indicador para el desarrollo de escalas. Aunque en el momento de construir una escala son mucho más importantes los argumentos de validez, una vez está asegurado que varios ítems son relevantes y representativos para medir un constructo, es posible seleccionar aquellos que producirán una escala con mayor consistencia interna. Tradicionalmente, la contribución de un ítem a la fiabilidad de la suma se ha evaluado utilizando un indicador conocido como "alfa sin el ítem" que consiste en valorar si la fiabilidad aumentaría o disminuiría al eliminar este ítem. Tampoco en este contexto consideramos adecuado utilizar un indicador basado en α, puesto que compartiría todos los problemas que hemos discutido en los apartados anteriores. Afortunadamente, el coeficiente omega puede ser utilizado para calcular la fiabilidad de consistencia interna de las puntuaciones obtenidas con cualquier subconjunto de ítems y, en particular, con todos los ítems excepto aquel cuya contribución al conjunto deseamos estudiar.

El procedimiento específico fue desarrollado por Raykov y sus colegas en tres artículos sucesivos. La propuesta inicial para ítems con una escala de respuesta cuantitativa (Raykov, 2007) fue generalizada posteriormente para datos dicotómicos (Raykov et al., 2010) y finalmente, a condiciones más generales, a saber, datos no normales, escalas multidimensionales, presencia de errores correlacionados o datos faltantes (Raykov y Marcoulides, 2016a). De nuevo, siguiendo su costumbre, los autores

incluyen apéndices que contienen la sintaxis necesaria en el paquete comercial de Mplus. A continuación se presenta un procedimiento inspirado en las ideas de los tres artículos.

Primero, se deberá fijar un valor de referencia para la fiabilidad deseada de la escala. Este puede provenir del conocimiento normativo (e.g., una consistencia interna mayor o igual a .70) o por estudios previos en el campo (e.g., para emular la consistencia interna de una escala publicada en otra cultura) o puede derivarse de datos obtenidos utilizando la escala en su estado actual de desarrollo. A continuación se ajustará el modelo de medida y si se desea se podrá obtener el coeficiente omega de la puntuación total. El siguiente paso consistirá en utilizar los parámetros estimados para calcular omega en el subconjunto formado por todos los ítems exceptuando el primero, y en replicar este cálculo para cada ítem, de manera que se obtendrá un indicador "omega sin el ítem" para cada uno. Estos indicadores podrían compararse visualmente con el valor de referencia elegido, análogamente al procedimiento habitual utilizado con "alfa sin el ítem". Sin embargo, Raykov y sus colegas proponen tomar decisiones basadas en el IC de la diferencia entre el coeficiente de referencia menos el coeficiente "omega sin el ítem". La contribución de un elemento a la consistencia interna es relevante si el IC de la diferencia no incluye el valor cero. Si el coeficiente de referencia fuera la fiabilidad actual de la escala, la interpretación sería la siguiente: en caso de que todos los valores incluidos en el IC sean superiores a cero, el ítem será útil, ya que si se eliminase, la escala perdería fiabilidad. En caso de que todos los valores incluidos en el IC sean inferiores a 0, entonces sería preferible excluir el ítem puesto que su presencia empeoraría la fiabilidad global.

A modo de síntesis: Regresando a casa con nuevas ideas para la obtención y el análisis de los datos

El objetivo de este trabajo ha sido facilitar la incorporación del conocimiento psicométrico más reciente respecto a la estimación de la fiabilidad de consistencia interna de las puntuaciones obtenidas a través de cuestionarios a la labor cotidiana de los investigadores y revisores en los ámbitos de las ciencias sociales y de la salud. Para ello, en primer lugar hemos examinado las razones para utilizar α o bien el coeficiente omega en la estimación de la fiabilidad de consistencia interna en escalas unidimensionales. Hemos proporcionado dos tipos de razones metodológicas para la toma de decisiones, unas basadas en el modelo de medida subyacente a los datos y otras en estudios de simulación sobre el sesgo que supone utilizar cualquiera de los dos coeficientes. En segundo lugar, hemos ofrecido una guía práctica para desarrollar el análisis, proporcionando en dos apéndices la sintaxis necesaria en el entorno de *software* libre R y comentando los resultados de varios ejemplos. Finalmente, hemos esbozado las ideas principales para la aplicación de los conceptos básicos al análisis de cuestionarios dimensionalmente complejos, a diseños multinivel con datos faltantes y al desarrollo de escalas. En este apartado final, extraeremos algunas consecuencias prácticas sobre el diseño y los procedimientos de obtención de datos, así como sobre su análisis, derivadas del razonamiento seguido a lo largo del artículo.

Al preparar el diseño de recogida de datos más usual basado en una única administración de una sola prueba, los investigadores toman decisiones que condicionarán definitivamente tanto el futuro análisis de datos como sus resultados. Destacamos tres de estas decisiones, que tienen que ver con la determinación del tamaño de la muestra, con la obtención de covariables predictoras de los datos faltantes y con la elaboración de procedimientos para obtener datos lo más completos posible y prestando atención a que los procesos de respuesta sean los esperados.

La determinación del tamaño de la muestra para la estimación de la fiabilidad debe ser puesta en contexto. Por una parte, la fiabilidad de las medidas suele estimarse en un estudio piloto con relativamente pocos casos de manera que el uso ingenuo de α puede proporcionar estimaciones muy sesgadas. Por otra parte, las estimaciones más correctas, basadas en los métodos SEM, requieren de grandes muestras para lograr resultados estables (Yang y Green, 2010), por lo que los costes del estudio piloto podrían aumentar desproporcionadamente debido a la adopción de esta metodología. Por lo tanto, antes de agradecer a α sus servicios y usar a partir de ahora los estimadores derivados de SEM (McNeish, 2017), o antes de evitar por completo el uso de modelos SEM para este cometido debido a sus dificultades (Davenport et al., 2016), vale la pena considerar cuándo necesitamos cambiar de alfa a omega. El conocimiento del cuestionario y de su desempeño psicométrico en estudios previos puede ayudar enormemente a limitar el coste del estudio piloto sin comprometer la estimación correcta de la fiabilidad. De acuerdo con los resultados de los estudios de simulación discutidos a lo largo de este trabajo, α es un buen estimador de la fiabilidad para modelos congenéricos con cargas factoriales altas y un gran número de ítems (Gu et al., 2013). La principal amenaza para una correcta estimación de la fiabilidad proviene de la presencia de errores correlacionados no corregidos o de efectos de método, una amenaza que empeora cuando el cociente entre las cargas factoriales y el error es bajo (Gu et al., 2013) y en medidas basadas en un número de ítems reducido (Graham, 2006).

En consecuencia, en el caso de que los datos psicométricos previos mostraran cargas factoriales altas para todos los ítems en un factor y sin efectos espurios, sería pertinente optar por una primera aproximación a la estimación de la fiabilidad de consistencia interna usando α. Se podría posponer un análisis más completo del modelo de medida hasta obtener los datos del estudio principal, que normalmente se basará en muestras más grandes que son más adecuadas para este propósito. La adopción de esta estrategia podría mantener el tamaño de la muestra y los costes del estudio piloto dentro de límites razonables.

Por el contrario, si el cuestionario contiene unos pocos ítems por factor, o si existen dudas acerca del tamaño de las cargas factoriales o de la unidimensionalidad, sería más seguro estimar la fiabilidad de consistencia interna partiendo del modelo de medida apropiado usando la metodología SEM y, en consecuencia, trabajando con muestras más grandes desde el principio. Por último, en caso de que los resultados previos pusieran en peligro la calidad de la medida, sería bueno disponer de dicha información en la etapa de diseño cuando se decidieran las medidas que se incluirían en el estudio principal, de manera que pudieran incluirse también en el estudio piloto.

En cuanto a los datos faltantes y al proceso de respuesta, aunque se han desarrollado métodos estadísticos robustos para

afrontar tanto los datos incompletos como los sesgos de respuesta, el mejor momento para abordarlos es durante la fase de recogida de datos. Deben hacerse todos los esfuerzos para facilitar la participación de los encuestados con el fin de aumentar la calidad de los datos y, en última instancia, de las conclusiones. Adicionalmente, se aconseja registrar los posibles predictores de datos faltantes. Como hemos visto (véase también Raykov y Marcoulides, 2016a), la inclusión de dichos predictores en análisis posteriores permitirá corregir el sesgo debido a datos faltantes que no se distribuyen aleatoriamente.

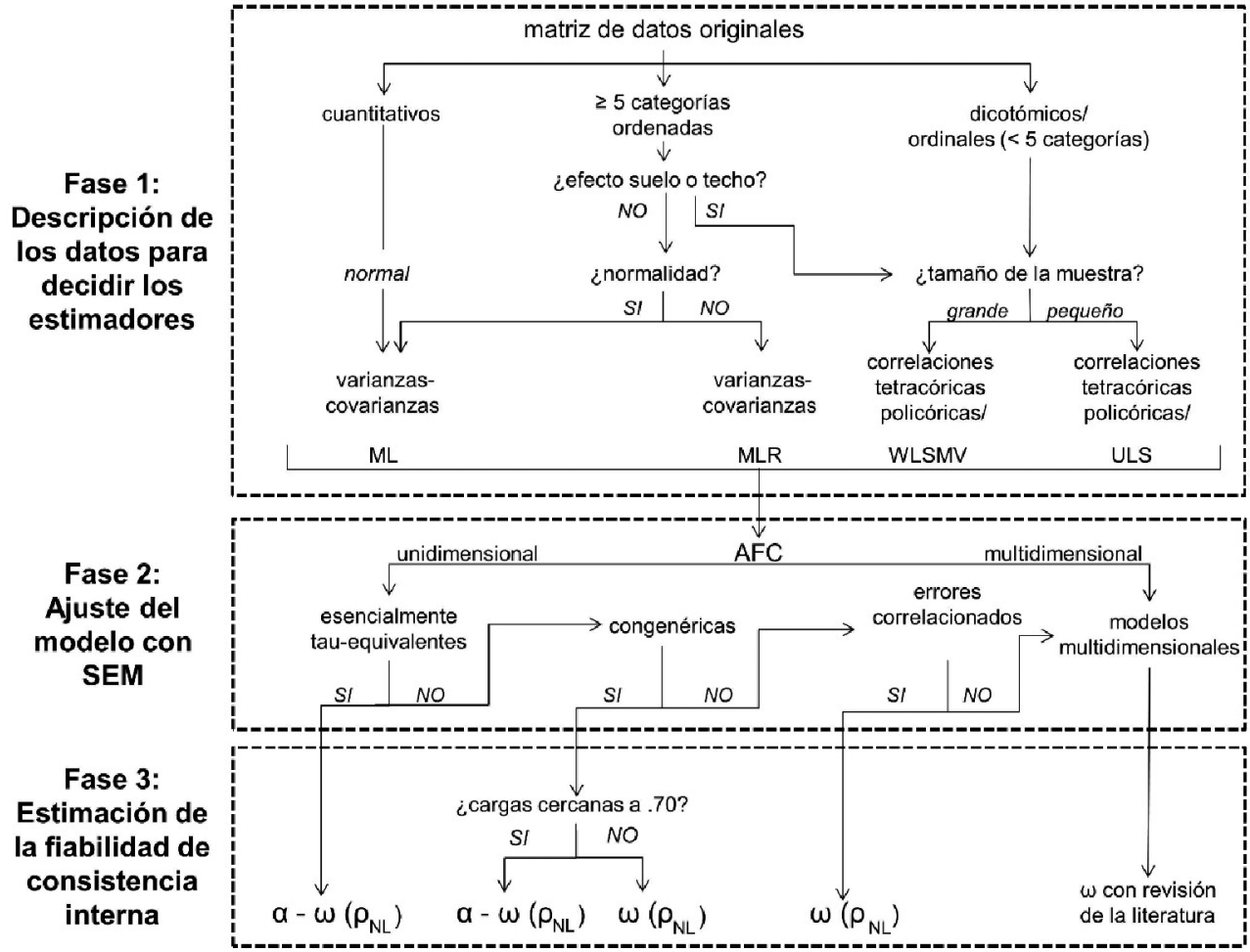

Figura 5. Diagrama de decisión de las tres fases analíticas implicadas en la estimación de la consistencia interna basada en el análisis factorial confirmatorio. *Nota.* Entre paréntesis el coeficiente aconsejado para datos tratados de forma ordinal. SEM = modelo de ecuaciones estructurales; AFC = análisis factorial confirmatorio; ML = Máxima verisimilitud; MLR = máxima verosimilitud robusta; WLSMV = mínimos cuadrados ponderados ajustados por media y varianza; ULS = mínimos cuadrados no ponderados; α = coeficiente alfa de Cronbach; ω = coeficiente(s) de fiabilidad basados en SEM lineal; ρ_{NL} = fiabilidad no lineal basada en SEM. Véanse los detalles en el texto.

Por lo que respecta al análisis de los datos, en la Figura 5 presentamos un diagrama de decisiones que sintetiza las ideas desarrolladas en los apartados anteriores. De acuerdo con el enfoque del artículo, proponemos realizar un análisis en tres fases. Durante la Fase 1, deben tenerse en cuenta el tamaño de la muestra, el formato de la escala de respuesta y también los resultados de la exploración de las distribuciones univariadas y de las relaciones entre los ítems. La primera decisión, basada en las distribuciones de las respuestas, sería si los datos se tratarán como cuantitativos o como ordinales/categóricos. Cuando la escala de respuesta es cuantitativa y las distribuciones normales, los parámetros del modelo de medida se estimarán por ML a partir de la matriz de varianzas-covarianzas. Cuando la escala de respuesta tiene cinco o más categorías y las distribuciones de respuestas no presentan acumulación de casos en los extremos de la escala, los datos pueden ser tratados como cuantitativos. En este caso, se analizará la matriz de varianzas-covarianzas con un estimador ML de forma general, corrigiendo las desviaciones menores de la normalidad mediante la estimación robusta de los errores estándar (MLR). Por el contrario, cuando el número de categorías en la escala de respuesta es menor de cinco, o cuando es igual o mayor que cinco pero se observan efectos claros de suelo o techo, los datos deben ser tratados como ordinales o categóricos. En este caso, se factorizará la matriz de correlaciones policóricas o tetracóricas utilizando en general el estimador

WLSMV, o el estimador ULS si las muestras son de tamaño pequeño.

En la Fase 2 del análisis se ajusta el modelo de medida. La decisión a tomar consiste en elegir el modelo de medida que muestre buen ajuste, sea más parsimonioso y tenga un contenido interpretable. De acuerdo con la línea argumental de este artículo, en la Figura 5 se representa el análisis empezando por el modelo más restrictivo que es el de medidas esencialmente tau-equivalentes y avanza relajando sucesivamente sus supuestos. Sin embargo, el análisis puede iniciarse comprobando el ajuste del modelo que sea más plausible de acuerdo con las expectativas de los investigadores y con la exploración inicial de las relaciones entre los ítems. Si la escala es al menos esencialmente unidimensional, la decisión involucra tres modelos anidados, el de medidas esencialmente tau equivalentes, el de medidas congenéricas y el de medidas con errores correlacionados o bifactor. Por otra parte, si ninguno de estos modelos encaja con los datos, o si la escala es multidimensional, se explorarán opciones de modelado más complejas, como los modelos con factores correlacionados, con factores de segundo orden, bifactor con factores de grupo no espurios, o de ítems con cargas factoriales cruzadas.

Por último, en la Fase 3, recomendamos que se elija el coeficiente de consistencia interna atendiendo al tipo de datos, a la estructura del modelo de medida y a la visión del investigador sobre la configuración de la varianza verdadera. Nos referimos en primer lugar los datos tratados como cuantitativos. Si el modelo de medidas esencialmente tau-equivalentes ha recibido apoyo empírico, la fiabilidad de consistencia interna de la suma o promedio de ítems podrá estimarse tanto con α como con el coeficiente omega. Si el que recibe apoyo es el modelo de medidas congenéricas, será más apropiado el uso de omega, aunque la diferencia con α no sería muy prominente en caso de que las cargas factoriales fuesen altas. Si se ha aceptado un modelo con errores correlacionados o un modelo bifactor con un factor general y otros espurios, entonces entra en juego la visión del investigador sobre las fuentes de variación verdadera. En caso que las fuentes espurias fueran tratadas como parte de la varianza del error, sería apropiado utilizar la fórmula corregida por errores correlacionados o utilizar omega jerárquico, mientras que si se considerasen parte de la varianza verdadera, se debería usar omega total. Finalmente, si se ha aceptado un modelo de medida multidimensional, para el cálculo de la fiabilidad es apropiado prestar atención a las recomendaciones de la literatura metodológica, puesto que en cada caso debe derivarse la fórmula correcta para el cálculo de las varianzas verdadera y observada en cada una de las subescalas de las que se desee estimar la fiabilidad.

Para datos ordinales unidimensionales, la mejor opción es utilizar el coeficiente de fiabilidad no lineal basado en SEM desarrollado por Green y Yang (2009), puesto que estima la varianza verdadera y observada en la métrica de la suma de los ítems. Este coeficiente se ha representado entre paréntesis en la Figura 5. De nuevo, y particularmente en los modelos con errores correlacionados, deberá prestarse especial atención a que las cargas factoriales estén correctamente estimadas con el fin de asegurar que la varianza verdadera que figura en el numerador de la fórmula sea correcta. Tal como se ha visto al discutir la Ecuación 3, el denominador no es tan dependiente del modelo puesto que la varianza observada siempre podría calcularse a partir de los datos observados. Esta fórmula para la fiabilidad no lineal es de desarrollo más reciente, de manera que es de esperar que se siga estudiando su rendimiento en diversas condiciones en el futuro.

También sería útil considerar qué otros coeficientes se deberían aportar en cada estudio. Por ejemplo, podría ser interesante informar el valor de α de forma rutinaria, y en caso de que α y el coeficiente derivado de SEM difirieran, comentar cuál de los estimadores es más creíble con base en el modelo de medida. Si este hábito estuviera extendido, podría ser útil, al menos, para dos finalidades. En primer lugar, cuando se aplicara a cuestionarios bien conocidos, proporcionaría nuevos datos psicométricos comparables a los de los estudios anteriores, donde muy probablemente sólo se incluyó el valor de α. En segundo lugar, y lo que es más importante, ayudaría a aumentar el conocimiento sobre el rendimiento de α en diversidad de contextos aplicados y, en particular, a evaluar en qué casos la diferencia entre α y omega sería prácticamente insignificante. Tal como sugieren Raykov y Marcoulides (2015), esto constituiría una contribución importante para salvar la brecha entre la literatura metodológica y la aplicada. Además, si el modelo de medida implica efectos de método, factores espurios o errores correlacionados, sería muy conveniente informar de los coeficientes omega jerárquico y omega total, así como discutir sus posibles diferencias como sugieren Green y Yang (2015), y también derivar las consecuencias esperadas sobre el desempeño de la medida en diversos contextos como el de la predicción, la comparación de grupos y los estudios longitudinales.

Otro aspecto importante a tener en cuenta es que cualquier estimación de la fiabilidad basada en los modelos SEM depende del modelo de medida subyacente, y también de varios aspectos del análisis. Los más importantes son (a) el método de estimación de parámetros utilizado (e.g., ML, MLR, ULS, WLSMV), (b) la fórmula específica que puede basarse en la matriz de covarianzas reproducida por el modelo como en la Ecuación 2 o en la matriz de covarianzas observada como en la Ecuación 3, y (c) las técnicas de estimación del IC, tales como varios tipos de métodos *bootstrap* o el método delta. Por lo tanto, se recomienda proporcionar toda esta información en cada estudio con el objetivo de facilitar su comprensión, su réplica y su inclusión correcta en los estudios meta-analíticos

Como se mencionó en la introducción, nuestra propuesta considera el coeficiente de fiabilidad de consistencia interna como un subproducto del modelo de medición. Nuestro punto de vista está de acuerdo con el de otros autores que sugieren siempre ajustar un modelo de medición y obtener los coeficientes de fiabilidad a partir de los parámetros estimados (e.g., Crutzen y Peters, 2015; Graham, 2006; Green y Yang, 2015).En este sentido, nuestra aportación consiste en destacar explícitamente la fase previa de análisis exploratorio de los datos. También estamos alineados con la idea de que es el momento de proporcionar recursos para que las prácticas correctas se integren en las rutinas de trabajo de investigadores y revisores como defienden por ejemplo Cho (2016); Dunn, Baguley, y Brunsden (2014) o Zhang y Yuan (2016). Sin embargo, no estamos totalmente de acuerdo con algunos de ellos en que proporcionando recursos simplificados para que el analista pueda estimar omega a través de unos pocos clics se conseguirán publicaciones con

mejores estimaciones de la fiabilidad. En nuestra opinión, la gama de modelos de medida y técnicas de estimación de parámetros a considerar dificulta, si no es que imposibilita, el desarrollo de un recurso simplificado global de este tipo. Por ejemplo, la calculadora Excel™ de Cho (2016) trata diversos modelos que se aplican a la suma de ítems cuantitativos; las reglas dadas por Dunn et al. (2014) para calcular omega usando R, son adecuadas para medidas unidimensionales y siempre que se desee calcular la fiabilidad derivada del AFC lineal; y la interfaz basada en R de Zhang y Yuan (2016) para estimar robustamente α y omega puede aplicarse a una restringida gama de modelos para datos cuantitativos. Se trata pues de recursos que pueden ser útiles en algunos casos, pero que resultan claramente insuficientes en la medida en que las escalas de respuesta utilizadas en cuestionarios sean ordinales.

Por el contrario, nosotros creemos que un analista bien informado es el mejor garante de un análisis correcto y, en última instancia, de publicaciones con mejores estimaciones de fiabilidad. Por lo tanto, una conclusión final es que deben hacerse todos los esfuerzos para desarrollar el análisis basado en un profundo conocimiento de la teoría y de los resultados anteriores relacionados con el cuestionario, así como en un amplio conocimiento de las posibilidades de diseño, modelado estadístico y estimaciones apropiadas de los coeficientes de fiabilidad de consistencia interna. Esto requiere de investigadores y revisores con formación especializada tanto en la esfera aplicada como en la metodológica y creemos que esta formación sería particularmente útil para cerrar la brecha entre los desarrollos metodológicos y las prácticas de investigación aplicada. Al compartir la sintaxis en el entorno de *software* libre R y al aplicarla a algunos ejemplos simples pero arquetípicos, confiamos en estimular la curiosidad de nuestros lectores para realizar un análisis completo a partir de los datos que proporcionamos, y también que se sientan tentados a aplicar todo el procedimiento a sus propias necesidades de estimación de fiabilidad de consistencia interna.

Agradecimientos.- Los autores de este trabajo agradecen las ayudas de la Dirección General de Investigación y Gestión del Plan Nacional de I+D+i, del Ministerio de Economía y Competitividad, a través de los proyectos EDU2013-41399-P y DEP2014-52481-C3-1-R y la ayuda de la Agencia de Gestión de Ayudas Universitarias y de Investigación AGAUR de la Generalitat de Catalunya (2014 SGR 224).

Referencias

Abad, F. J., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en ciencias sociales y de la salud* [*Measurement in social and health sciences*]. Madrid: Síntesis.

American Psychological Association (2010). *Publication manual of the American Psychological Association.* (6th ed.). Washington, DC.

Behrens, J. T., DiCerbo, K. E., Yel, N., & Levy, R. (2012). Exploratory Data Analysis. In *Handbook of Psychology, Second Edition*. John Wiley & Sons, Inc. doi:10.1002/9781118133880.hop202002

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*(1), 137–143. doi:10.1007/s11336-008-9100-1

Bentler, P. M. (2016). Specificity-enhanced reliability coefficients. *Psychological Methods*, *0*. Advance online publication. doi:10.1037/met0000092

Birnbaum, A. (1968). Some latent trait models and their use in inferring a examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Black, R. A., Yang, Y., Beitra, D., & McCaffrey, S. (2015). Comparing fit and reliability estimates of a psychological instrument using second-order CFA, bifactor, and essentially tau-equivalent (coefficient alpha) Models via AMOS 22. *Journal of Psychoeducational Assessment*, *33*(5), 451–472. doi:10.1177/0734282914553551

Bovaird, J. A., & Koziol, N. A. (2012). Measurement models for ordered-categorical indicators. In *Handbook of Structural Equation Modeling* (pp. 495–511). New York, NY: The Guilford Press.

Bollen, K. (1980). Issues in the comparative measurement of political democracy. *American Sociological Review*, *45*, 370–390.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research.* 2nd Ed. London: The Guilford Press.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*(4), 796–846. doi:10.1111/j.1467-6494.2011.00749.x

Cheng, Y., Yuan, K. H., & Liu, C. (2012). Comparison of reliability measures under factor analysis and item response theory. *Educational and Psychological Measurement*, *72*(1), 52–67. doi:10.1177/0013164411407315

Cho, E. (2016). Making Reliability Reliable: A Systematic approach to reliability coefficients. *Organizational Research Methods*, *19*(4), 651–682. doi:10.1177/1094428116656239

Conway, J. M., & Lance, C. E. (2010). What reviewers should expect from authors regarding common method bias in organizational research. *Journal of Business and Psychology*, *25*(3), 325–334. doi:10.1007/s10869-010-9181-6

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391–418. doi:10.1177/0013164404266386

Crutzen, R., & Peters, G. (2015). Scale quality: alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychology Review*, 1–6. doi:10.1080/17437199.2015.1124240

Davenport, E. C., Davison, M. L., Liou, P. Y., & Love, Q. U. (2016). Easier said than done: rejoinder on Sijtsma and on Green and Yang. *Educational Measurement: Issues and Practice*, *35*(1), 6–10. doi:10.1111/emip.12106

Deng, L., & Chan, W. (2016). Testing the difference between reliability coefficients alpha and omega. *Educational and Psychological Measurement*, *online*, 1–19. doi:10.1177/0013164416658325

Dimitrov, D. M. (2003). Reliability and true-score measures of binary items as a function of their Rasch difficulty parameter. *Journal of Applied Measurement*, *4*(3), 222–233. doi:10.1177/0146621603258786

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. doi:10.1111/bjop.12046

Elosua, P., & Zumbo, B. D. (2008). Reliability coefficients for ordinal response scales. *Psicothema*, *20*(4), 896–901.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: The Guilford Press.

Enders, C. K. (2013). Dealing with missing data in developmental research. *Child Development Perspectives*, *7*(1), 27–31. doi:10.1111/cdep.12008

Ferrando, P. J., & Lorenzo-Seva, U. (2014). El análisis factorial exploratorio de los ítems: Algunas consideraciones adicionales [Exploratory item factor analysis: some additional considerations] *Anales de Psicología*, *30*(3), 1170–1175. doi:10.6018/analesps.30.3.199991

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: origins, development and future directions. *Psicothema*, *29*(2), 236–240. https://doi.org/10.7334/psicothema2016.304

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, *16*(4), 625–641. doi:10.1080/10705510903203573

Gabler, S., & Raykov, T. (2017). Evaluation of maximal reliability for unidimensional measuring instruments with correlated errors. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(1), 104-111. Advance online publication. doi:10.1080/10705511.2016.1159916

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research and Evaluation*, *17*(3), 1–13.

Gignac, G. E. (2014). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment*, *30*(2), 130–139. doi:10.1027/1015-5759/a000181

Graham, J. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. What they are and how to use them. *Educational and Psychological Measurement*, *66*(6), 930–944. doi: 10.1177/0013164406288165.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–76. doi:10.1146/annurev.psych.58.110405.085530

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*(1), 155–167. doi:10.1007/s11336-008-9099-3

Green, S. B., & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: coefficient alpha and omega coefficients. *Educational Measurement: Issues and Practice*, *34*(4), 14–20. doi:10.1111/emip.12100

Green, S. B., Yang, Y., Alt, M., Brinkley, S., Gray, S., Hogan, T., & Cowan, N. (2016). Use of internal consistency coefficients for estimating reliability of experimental task scores. *Psychonomic Bulletin & Review*, *23*(3), 750–763. doi:10.3758/s13423-015-0968-3

Gu, F., Little, T. D., & Kingston, N. M. (2013). Misestimation of reliability using coefficient alpha and structural equation modeling when assumptions of tau-equivalence and uncorrelated errors are violated. *Methodology*, *9*(1), 30–40. doi:10.1027/1614-2241/a000052

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.), *Structural Equation Modeling: Present and future* (pp. 195–216). Lincolnwood, IL: Scientific Software International, Inc.

Hancock, G. R., & Mueller, R. O. (Eds.). (2013). *Structural equation modeling. A second course* (2nd ed.). Charlotte, NC: Information Age Publishing.

Hoyle, R. H. (Ed.). (2012). *Handbook of Structural equation modeling*. New York, NY: The Guilford Press.

Huggins-Manley, A. C., & Han, H. (2017). Assessing the sensitivity of weighted least squares model fit indexes to local dependence in item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(3), 331–340. doi:10.1080/10705511.2016.1247355

Izquierdo, I., Olea, J., & Abad, F. J. (2014). Exploratory factor analysis in validation studies: Uses and recommendations. *Psicothema*, *26*(3), 395–400. doi:10.7334/psicothema2013.349

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133. doi:10.1007/s00170-004-2446-3

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients : Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods*, *21*(1), 69–92. doi:10.1037/a0040086

Kim, S., & Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review*, *11*(2), 179–188. doi:10.1007/s12564-009-9062-8

Lance, C. E., Dawson, B., Birkelbach, D., & Hoffman, B. J. (2010). Method effects, measurement error, and substantive conclusions. *Organizational Research Methods*, *13*(3), 435–455. doi:10.1177/1094428109352528

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*(3), 285–300. doi:10.1037/a0033266

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. [Exploratory item factor analysis: A practical guide revised and up-dated] *Anales de Psicología*, *30*(3), 1151–1169. doi:10.6018/analesps.30.3.199361

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Malone, P. S., & Lubansky, J. B. (2012). Preparing data for structural equation modeling: doing your homework. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 263–276). New York, NY: The Guilford Press.

Mansolf, M., & Reise, S. P. (2016). Exploratory bifactor analysis: The Schmid-Leiman orthogonalization and Jennrich-Bentler analytic rotations. *Multivariate Behavioral Research*, *51*(5), 698–717. doi:10.1080/00273171.2016.1215898

Marsh, H. W. (1996). Positive and negative global self-esteem: a substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*(4), 810–819. doi:10.1037/0022-3514.70.4.810

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*(3), 471–91. doi:10.1037/a0019227

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right--camouflaging misspecification with item parcels in CFA models. *Psychological Methods*, *18*, 257–84. doi:10.1037/a0032773

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*(4), 344–362. doi:10.1037/1082-989X.11.4.344

Maydeu-Olivares, A., Fairchild, A. J., & Hall, A. G. (2017). Goodness of fit in item factor analysis: effect of the number of response alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *0*(0), 1–11. doi:10.1080/10705511.2017.1289816

McCrae, R. R. (2014). A more nuanced view of reliability: specificity in the trait hierarchy. *Personality and Social Psychology Review*, *19*(2), 97–112. doi:10.1177/1088868314541857

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

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 0*. Advance online publication. doi: 10.1037/met0000144

Muñiz, J. (1992). *Teoría clásica de los tests* [*Classical test theory*]. Madrid: Pirámide.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide (Eighth Edition)*. Los Angeles, CA: Muthén & Muthén.

Muthén & Muthén (n.d.). *Chi-Square difference testing using the Satorra-Bentler scaled Chi-Square*. Retrieved June 19, 2017 from https://www.statmodel.com/chidiff.shtml

Napolitano, C. M., Callina, K. S., & Mueller, M. K. (2013). Comparing alternate approaches to calculating reliability for dichotomous data: The sample case of adolescent selection, optimization, and compensation. *Applied Developmental Science*, *17*(3), 148–151. doi:10.1080/10888691.2013.804372

Ntoumanis, N., Mouratidis, T., Ng, J. Y. Y., & Viladrich, C. (2015). Advances in quantitative analyses and their implications for sport and exercise psychology research. In S. Hanton & S. D. Mellalieu (Eds.), *Contemporary advances in sport psychology: A review*. (pp. 226–257). London: Routledge.

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGrawHill.

Padilla, M. A., & Divers, J. (2016). A comparison of composite reliability estimators: coefficient omega confidence intervals in the current literature. *Educational and Psychological Measurement*, *76*(3), 436–453. doi:10.1177/0013164415593776

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrived from https://www.R-project.org/.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*(2), 173–184. doi: 0803973233

Raykov, T. (1998). Coefficient alpha and composite reliability with

interrelated nonhomogeneous items. *Applied Psychological Measurement*, *22*(4), 375–385. doi:10.1177/014662169802200407

Raykov, T. (2001). Bias of coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, *25*(1), 69–76. doi:10.1177/01466216010251005

Raykov, T. (2004). Point and interval estimation of reliability for multiple-component measuring instruments via linear constraint covariance structure modeling. *Structural Equation Modeling*, *11*(3), 452–483. doi:10.1207/s15328007sem1103

Raykov, T. (2007). Reliability if deleted, not "alpha if deleted": Evaluation of scale reliability following component deletion. *The British Journal of Mathematical and Statistical Psychology*, *60*(2), 201–216. doi: 10.1348/000711006X115954

Raykov, T. (2012). Scale construction and development using structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 472–492). New York, NY: Guilford Press.

Raykov, T., Dimitrov, D. M., & Asparouhov, T. (2010). Evaluation of scale reliability with binary measures using latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(2), 265–279. doi:10.1080/10705511003659417

Raykov, T., Gabler, S., & Dimitrov, D. M. (2016). Maximal reliability and composite reliability: examining their difference for multicomponent measuring instruments using latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 384–391. doi:10.1080/10705511.2014.966369

Raykov, T., & Marcoulides, G. A. (2012). Evaluation of validity and reliability for hierarchical scales using latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(3), 495–508. doi:10.1080/10705511.2012.687675

Raykov, T., & Marcoulides, G. A. (2014). Scale reliability evaluation with heterogeneous populations. *Educational and Psychological Measurement*, *75*(5), 875–892. doi:10.1177/0013164414558587

Raykov, T., & Marcoulides, G. A. (2015). A direct latent variable modeling based method for point and interval estimation of coefficient alpha. *Educational and Psychological Measurement*, *75*(1), 146–156. doi:10.1177/0013164414526039

Raykov, T., & Marcoulides, G. A. (2016a). Scale reliability evaluation under multiple assumption violations. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(2), 302–313. doi:10.1080/10705511.2014.938597

Raykov, T., & Marcoulides, G. A. (2016b). On Examining specificity in latent construct indicators. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 845-855. doi:10.1080/10705511.2016.1175947

Raykov, T., & Pohl, S. (2013). Essential unidimensionality examination for multicomponent scales: an interrelationship decomposition approach. *Educational and Psychological Measurement*, *73*(4), 581–600. doi:10.1177/0013164412470451

Raykov, T., West, B. T., & Traynor, A. (2015). Evaluation of coefficient alpha for multiple-component measuring instruments in complex sample designs. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 429–438. doi:10.1080/10705511.2014.936081

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. doi:10.1080/00273171.2012.715555

Revelle, W. (2016). *psych: Procedures for personality and psychological research*. R package version 1.6.4. North-western University, Evanston. Retrieved June 19, 2017 from http://cran.r-project.org/web/packages/psych/.

Revelle, W., & Zinbarg, R. E. (2009). Coefficient alpha, beta, omega, and the GLB: Comment on Sitjsma. *Psychometrika*, *74*(1), 145–154.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. doi:10.1037/a0029315

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. doi:10.1037/met0000045

Rosseel, Y. (2012). lavaan: An R Package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Sass, D. a., Schmitt, T. a., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 167–180. doi:10.1080/10705511.2014.882658

semTools Contributors. (2016). semTools: Useful tools for structural equation modeling. R package version 0-4-11.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. doi:10.1007/s11336-008-9101-0

Sijtsma, K. (2015). Delimiting coefficient alpha from internal consistency and unidimensionality. *Educational Measurement: Issues and Practice*, *34*(4), 10–13.

Spector, P. E. (2006). Method variance in organiztional research. Truth or urban legend ? *Organizational Research Methods*, *9*(2), 221–232. doi:1094428105284955

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589–617. doi:10.1007/BF02294821

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12), 1-20. Retrieved from http://www.jstatsoft.org/v21/i12/.

Yang, Y., & Green, S. B. (2010). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(1), 66–81. doi:10.1080/10705510903438963

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st Century? *Journal of Psychoeducational Assessment*, *29*(4), 377–392. doi:10.1177/0734282911406668

Yang, Y., & Green, S. B. (2015). Evaluation of structural equation modeling estimates of reliability for scales with ordered categorical items. *Methodology*, *11*(1), 23–34. doi:10.1027/1614-2241/a000087

Zhang, Z., & Yuan, K.-H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. *Educational and Psychological Measurement*, *76*(3), 387–411. doi:10.1177/0013164415594658

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ωH: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133. doi:10.1007/s11336-003-0974-7

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods*, *6*(1), 21–29. doi:10.1107/S0907444909031205

(Artículo recibido: 22-09-2016; revisado: 11-11-2016; aceptado: 16-05-2017)

Apéndice A

Sintaxis R utilizada para estimar la consistencia interna en cuatro escenarios prácticos.

Las bases de datos están disponibles en http://ddd.uab.cat/record/173917, en la Tabla 1 y en la Tabla 2 se muestra una selección de resultados y en el texto principal pueden consultarse los demás detalles. Se recomienda el Apéndice A para usuarios experimentados en R. Los principiantes en este entorno pueden encontrar útiles los comentarios del Apéndice B.

```
#Definición del directorio de trabajo
setwd("c:/directoriodetrabajo")

#Instalación de los paquetes necesarios para realizar los análisis
#¡No ejecutar si ya están instalados!
install.packages("reshape2", dependencies = TRUE)
install.packages("psych", dependencies = TRUE)
install.packages("lavaan", dependencies = TRUE)
install.packages("semTools", dependencies = TRUE)
install.packages("MBESS", dependencies = TRUE)

#Carga de los paquetes necesarios para realizar los análisis
#Ejecutar al comienzo de una nueva sesión de trabajo
library(reshape2)
library(psych)
library(lavaan)
library(semTools)
library(MBESS)

#Caso 1: medidas esencialmente tau-equivalente
#Lectura de los datos, ver la estructura de los datos en Tabla B1
C1<-read.table('Case1.txt',header=TRUE)

#Fase 1
#Porcentajes de respuestas
prop.table(table(melt(C1)),1)*100
#Otros estadísticos univariados
describeBy(C1)
#Correlaciones de Pearson
lowerCor(C1, digits = 3)

#Fase 2
#Especificación del modelo de medidas esencialmente tau-equivalentes
C1tau <- 'Factor1 =~ L*Y1 + L*Y2 + L*Y3 + L*Y4 + L*Y5 + L*Y6'
#Estimación y ajuste del modelo
CFA_C1tau <- cfa(C1tau, C1,std.lv = TRUE)
#Listado de resultados
summary(CFA_C1tau, fit.measures=TRUE)
#Especificación del modelo de medidas congenéricas
C1cong <- 'Factor1 =~ Y1 + Y2 + Y3 + Y4 + Y5 + Y6'
CFA_C1cong <- cfa(C1cong, C1,std.lv = TRUE)
summary(CFA_C1cong, fit.measures=TRUE)

#Fase 3
#Estimación puntual de los coeficientes alfa y omega
reliability(CFA_C1tau)
#Estimación por intervalo del coeficiente alfa
ci.reliability(data=C1, type='alpha', interval.type='bsil', B=500)
#Estimación por intervalo del coeficiente omega para el modelo de medidas esencialmente tau-equivalentes
ci.reliability(data=C1, type='alpha-CFA', interval.type='bsil', B=500)
```

```
#Caso 2: medidas congenéricas con cargas factoriales altas y homogéneas
#Lectura de los datos
C2<-read.table('Case2.txt',header=TRUE)

#Fase 1
#Porcentaje de respuestas
prop.table(table(melt(C2)),1)*100
#Otros estadísticos univariados
describeBy(C2)
#Correlaciones de Pearson
lowerCor(C2, digits = 3)

#Fase 2
C2tau <- 'Factor1 =~ L*Y1 + L*Y2 + L*Y3 + L*Y4 + L*Y5 + L*Y6'
#Estimación y ajuste del modelo
CFA_C2tau <- cfa(C2tau, C2,std.lv = TRUE)
#Listado de resultados
summary(CFA_C2tau, fit.measures=TRUE)
#Especificación, estimación y ajuste del modelo de medidas congenéricas
C2cong <- 'Factor1 =~ Y1 + Y2 + Y3 + Y4 + Y5 + Y6'
CFA_C2cong <- cfa(C2cong, C2,std.lv = TRUE)
summary(CFA_C2cong, fit.measures=TRUE)

#Fase 3
#Estimación puntual de los coeficientes alfa y omega
reliability(CFA_C2cong)
#Estimación por intervalo del coeficiente alfa
ci.reliability(data=C2, type='alpha', interval.type='bsil', B=500)
#Estimación por intervalo del coeficiente omega para medidas congenéricas
ci.reliability(data=C2, type='omega', interval.type='bsil', B=500)

#Caso 3: medidas con errores correlacionados
#Lectura de los datos
C3<-read.table('Case3.txt',header=TRUE)

#Fase 1
#Porcentaje de respuestas
prop.table(table(melt(C3)),1)*100
#Otros estadísticos univariados
describeBy(C3)
#Correlaciones de Pearson
lowerCor(C3, digits = 3)

#Fase 2
#Especificación, estimación y ajuste de los modelos de medidas esencialmente tau-equivalentes y modelos congenéricos
C3tau <- 'Factor1 =~ L*Y1 + L*Y2 + L*Y3 + L*Y4 + L*Y5 + L*Y6'
CFA_C3tau <- CFA(C3tau, C3, std.lv = TRUE)
summary(CFA_C3tau, fit.measures=TRUE)
C3cong <- 'Factor1 =~ Y1 + Y2 + Y3 + Y4 + Y5 + Y6'
CFA_C3cong <- cfa(C3cong, C3, std.lv = TRUE)
summary(CFA_C3cong, fit.measures=TRUE)
# Especificación, estimación y ajuste de los modelos de medidas con errores correlacionados
C3err_corr <- 'Factor1 =~ Y1 + Y2 + Y3 + Y4 + Y5 + Y6
   Y4 ~~ Y5
   Y4 ~~ Y6
   Y5 ~~ Y6'
CFA_C3err_corr <- cfa(C3err_corr, C3, std.lv = TRUE)
summary(CFA_C3err_corr, fit.measures=TRUE)
```

```
#Fase 3
#Estimación puntual de los coeficientes alfa y omega
reliability(CFA_C3err_corr)
#Estimación por intervalo no disponible

#Caso 4: datos categóricos ordenados
#Lectura de los datos
C4<-read.table('Case4.txt',header=TRUE)

#Fase 1
#Porcentaje de respuestas
prop.table(table(melt(C4)),1)*100
#Otros estadísticos univariados
describeBy(C4)
#Correlaciones policóricas
polychoric(C4)

#Fase 2
#Especificación, estimación y ajuste de los modelos de medidas esencialmente tau-eauivalentes y modelos
congenéricos para ítems con categorías ordenadas
C4tau <- 'Factor1 =~ L*Y1 + L*Y2 + L*Y3 + L*Y4 + L*Y5 + L*Y6'
CFA_C4tau <- cfa(C4tau, C4,std.lv = TRUE, ordered = names(C4))
summary(CFA_C4tau, fit.measures=TRUE)
C4cong <- ' Factor1 =~ Y1 + Y2 + Y3 + Y4 + Y5 + Y6'
CFA_C4cong <- cfa(C4cong, C4,std.lv = TRUE, ordered=names(C4))
summary(CFA_C4cong, fit.measures=TRUE)

#Fase 3
#Estimación puntual de los coeficientes alfa y omega
reliability(CFA_C4cong)
#Estimación por intervalo del coeficiente omega para ítems categóricos congenéricos
ci.reliability(data=C4,type='categorical', interval.type='bca')
```

Apéndice B

Guía para la estimación de la consistencia interna en cuatro escenarios utilizando R.

Se recomienda esta guía para principiantes en el entorno R. Las bases de datos están disponibles en http://ddd.uab.cat/record/173917, en la Tabla 1 y en la Tabla 2 se muestra una selección de resultados y en el texto principal pueden consultarse otros detalles.

Para estimar la fiabilidad de consistencia interna en R, (a) preparar R para una sesión de trabajo (b) leer los datos a analizar y (c) realizar el análisis con las tres fases recomendadas en el texto principal. En el Apéndice B se describen las principales características de R y las líneas de sintaxis necesarias para obtener los resultados en las Tablas 1 y 2. La forma más sencilla de ejecutar un ejemplo es pegar las líneas de sintaxis proporcionadas en el Apéndice A en R y, si es necesario, adaptarlas a su propio análisis.

Trabajo con R

Si no dispone del software libre R en su ordenador, puede descargarlo gratis en https://www.r-project.org/. La sintaxis proporcionada en este Apéndice puede utilizarse en cualquier interfaz de R, ya sea la sencilla Rconsole o las interfaces más desarrolladas como RCommander, RStudio o DeduceR, que proporcionan facilidades adicionales a la consola.

Para obtener resultados, ejecute R, espere a que aparezca el indicador > en la consola, escriba una línea de sintaxis junto al indicador, presione la tecla Enter y lea la salida debajo de la línea de sintaxis. Vea a continuación una descripción extremadamente sintética del lenguaje R, características de sintaxis, entrada / salida de archivos y comandos de instalación. Puede encontrar más información se en el sitio web de R (https://www.r-project.org/).

De acuerdo con el lenguaje R, puede usar funciones que leen los datos y crean objetos que se muestran como resultados. Por ejemplo, cuando se aplica a datos cuantitativos, la función reliability() produce un objeto que contiene α expresado en la Ecuación 3 y el coeficiente omega de la Ecuación 5. Todas las funciones de R están incluidas en paquetes. Por ejemplo, el paquete psych permite calcular los coeficientes de correlación de Pearson con la función lowerCor(), los coeficientes de correlación policórica con la función polychoric()y los coeficientes de fiabilidad con la función reliability(). Algunos paquetes están disponibles por defecto, pero muchos otros deben ser instalados y cargados antes de ser usados. Finalmente, R dispone de muchos paquetes y funciones que realizan el análisis (e.g., el IC para α puede obtenerse usando el paquete psych o el paquete MBESS). En la sintaxis proporcionada en el Apéndice A hemos hecho selección de algunos de ellos.

Volviendo a cuestiones relacionadas con la sintaxis, R es sensible a las mayúsculas, de manera que reliability(C1) no es lo mismo que reliability(c1) o que Reliability(C1). Entre los símbolos especiales que se encuentran en la sintaxis proporcionada, # denota un comentario que no será tenido en cuenta por R pero que puede ser de utilidad para los lectores humanos, < - se utiliza para almacenar un resultado en un nuevo objeto, + - * / = son los operadores matemáticos y lógicos obvios, y =~ se usa para definir factores en un AFC. En cuanto a las convenciones de nomenclatura, R es un poco flexible. Algunos nombres se escriben con tipografía camel (e.g., semTools), otros están separados por puntos (e.g., install.packages) o utilizan subrayados (e.g., CFA_C1tau).

Respecto a los archivos de entrada y salida, resulta útil utilizar un directorio de trabajo definido por el usuario. Los archivos de datos deben estar disponibles en el directorio de trabajo para ser leídos usando la sintaxis proporcionada en el Apéndice A.

Para preparar su sesión de trabajo, configure el directorio de trabajo y active todos los paquetes necesarios. El directorio de trabajo actual se encuentra con la función:

```
getwd()
```

Para cambiar el directorio de trabajo debe usarse la función setwd()indicando, entrecomillado, el nuevo directorio dentro de un paréntesis. Observe que en R, los directorios se definen con la barra inclinada hacia la derecha / en lugar de la barra inclinada hacia la izquierda \ que es habitual en otros entornos. Por ejemplo:

```
setwd("c:/workingdirectory")
```

La instalación de los paquetes se realiza con la función install.packages()indicando, entrecomillado y dentro del paréntesis, el nombre del paquete.

Para obtener los resultados de la Tabla 1, hemos usado los siguientes paquetes: reshape2 para reorganizar los datos con el fin de obtener las tablas de frecuencias o de proporciones psych para obtener los estadísticos univariables y las correlaciones de Pearson o policóricas. En cuanto a los resultados de la Tabla 2, se obtuvieron utilizando lavaan para especificar, estimar y ajustar todos los modelos de medida, semTools para calcular las estimaciones puntuales de los coeficientes omega y α, y MBESS para el cálculo de los intervalos de confianza. Por lo tanto, la sintaxis es la siguiente:

```
install.packages("reshape2", dependencies = TRUE)
install.packages("psych", dependencies = TRUE)
install.packages("lavaan", dependencies = TRUE)
install.packages("semTools", dependencies = TRUE)
install.packages("MBESS", dependencies = TRUE)
```

Al ejecutar estos comandos, se mostrará una lista de repositorios (mirror CRAN). Seleccione uno, preferiblemente cercano geográficamente, y espere a que aparezca el indicador > en la consola cuando finalice la instalación. Una vez instalados, los paquetes permanecerán en sus archivos locales de R hasta que se quieran eliminar usando la función remove.packages()con las mismas convenciones.

Cada vez que se inicie una nueva sesión de trabajo, deberá cargar los paquetes con la función library() indicando el nombre del paquete entre paréntesis:

```
library(reshape2)
library(psych)
library(lavaan)
library(semTools)
library(MBESS)
```

Desde ese momento hasta el final de la sesión de trabajo, todas las funciones de los paquetes cargados estarán disponibles. Si desea obtener información sobre un paquete en particular, por ejemplo, cómo definir una función correctamente, simplemente deberá escribir los símbolos ?? seguidos por el nombre del paquete:

```
?? semTools
```

Lectura de los datos

Se pueden leer datos de diferentes formatos, pero sugerimos el uso de un archivo de texto simple. La Tabla B1 muestra las primeras líneas del archivo de datos del Caso 1. Cada línea contiene datos de una persona a todos los ítems y cada columna contiene todas las respuestas a uno de los seis ítems. Los valores están separados por tabuladores. La primera fila del archivo contiene el nombre de cada ítem, en este caso Y1, Y2, Y3, Y4, Y5 e Y6. Este archivo se ha guardado con el nombre de Case1.txt. Durante la sesión de análisis, el archivo de datos debe estar disponible en el directorio definido como directorio de trabajo en R.

Tabla B1. Primeros cinco registros del Caso 1

Y1	Y2	Y3	Y4	Y5	Y6
2	2	3	3	2	1
3	4	2	3	3	4
4	4	3	4	4	3
3	2	4	3	3	3
1	3	2	3	3	2

Este tipo de archivo de datos puede leerse con la función read_table (). Dentro del paréntesis hay que incluir, al menos dos informaciones, el nombre del archivo de datos entre comillas, y a continuación si la primera fila contiene (header=TRUE) o no (header=FALSE) el nombre de las variables. En nuestra sintaxis la instrucción fue la siguiente:

```
C1<-read.table('Case1.txt', header=TRUE)
```

El contenido del archivo se transfiere, a través del símbolo <-, a un objeto cuyo nombre es especificado por el usuario (en este ejemplo C1) para referencias futuras. Para comprobar si la tabla ha sido definida correctamente, simplemente hay que escribir el nombre del objeto y pulsar la tecla Enter

```
C1
```

Ejecución del análisis

Caso 1: Análisis de medidas esencialmente tau-equivalentes

A continuación, se describe la sintaxis R necesaria para realizar las tres fases del análisis y conseguir los resultados incluidos en la Tabla 1 (Fase 1) y Tabla 2 (Fase 2 y Fase 3).

Fase 1: Descripción de los datos

La siguiente instrucción proporciona la tabla de frecuencias de respuesta a cada categoría de cada ítem:

```
table(melt(C1))
```

Utilizamos la siguiente instrucción para obtener la table de porcentajes:

```
prop.table(table(melt(C1)),1)*100
```

Los estadísticos descriptivos básicos, como la media, desviación estándar, asimetría y curtosis, se obtienen de la siguiente manera:

```
describeBy(C1)
```

La matriz de correlaciones de Pearson se obtiene con la instrucción:

```
lowerCor(C1, digits = 3)
```

A partir de los resultados de la Tabla 1, concluimos que las respuestas a los ítems en el Caso 1 pueden ser tratadas como cuantitativas, utilizando, por tanto, la estimación ML para comprobar los modelos. Para una discusión más detallada, ver el texto principal.

Fase 2: Determinación del modelo de medida con mejor ajuste

El modelo de medidas esencialmente tau-equivalente se ha definido de la siguiente forma:

```
C1tau <- 'Factor1 =~ L*Y1 + L*Y2 + L*Y3 + L*Y4 + L*Y5 + L*Y6'
```

Donde la variable latente (Factor1) se define (= ~) como la suma ponderada de los seis ítems (Y_i). El peso(L) es una constante para todos los ítems, especificando así el supuesto de tau-equivalencia esencial. El resultado se transfiere (<-) a un objeto que el usuario ha denominado C1tau.

La instrucción:

```
CFA_C1tau <- cfa(C1tau, C1,std.lv = TRUE)
```

Ejecuta un análisis factorial confirmatorio (cfa) bajo el modelo definido en C1tau sobre los datos almacenados en C1. Los resultados se estandarizan (std.lv = TRUE) y se almacenan (<-) en el objeto CFA_C1tau. El método de estimación por defecto es ML.

Los índices de bondad de ajuste del modelo se obtienen en forma de resumen del objeto CFA_C1tau con la siguiente instrucción:

```
summary(CFA_C1tau, fit.measures=TRUE)
```

El modelo de medidas congenéricas se define y ajusta de manera análoga, simplemente eliminando la restricción de pesos constantes y almacenando los resultados en un objeto con un nuevo nombre definido por el usuario (C1cong):

```
C1cong <- 'Factor1 =~ Y1 + Y2 + Y3 + Y4 + Y5 + Y6'
CFA_C1cong <- cfa(C1cong, C1,std.lv = TRUE)
summary(CFA_C1cong, fit.measures=TRUE)
```

Tal como era de esperar, los índices de bondad de ajuste de ambos modelos (Tabla 2) favorecen el modelo de medidas esencialmente tau-equivalentes, tal y como se discute con más detalle en el texto principal.

Finalmente, puede ser útil tener en cuenta que si se desea usar un estimador diferente al de defecto, se debe especificar el estimador deseado entre comillas. Por ejemplo, si se desease utilizar un estimador robusto para datos cuantitativos, la instrucción debería ser la siguiente

```
CFA_C1cong <- CFA(C1cong, C1,std.lv = TRUE, estimator ="MLR")
```

Fase 3: Obtención de los coeficientes de fiabilidad para medidas esencialmente tau-equivalentes

El listado de resultados de la instrucción

```
reliability(CFA_C1tau)
```

proporciona varios estimadores puntuales de la fiabilidad. En la Tabla 2 se incluyen dos como resultados para el Caso 1. El primero es α, etiquetado en el listado como *alpha,* y calculado a partir de la Ecuación 4 con un estimador ULS. El tercero del listado de resultados, etiquetado como *omega2,* se obtiene a partir de la fórmula general del coeficiente omega de la Ecuación 5. Cuando se aplica a medidas congenéricas o esencialmente tau-equivalentes, como en el Caso 1, el resultado es equivalente al que se obtiene con la Ecuación 2 debido al hecho de que todas las correlaciones entre los errores son iguales a cero.

Las instrucciones

```
ci.reliability(data=C1, type='alpha', interval.type='bsil', B=500)
ci.reliability(data=C1, type='alpha-CFA', interval.type='bsil', B=500)
```

proporcionan, respectivamente, los IC estimados para los coeficientes α y omega bajo el supuesto de medidas esencialmente tau-equivalentes. En primer lugar, se utiliza data= para indicar los datos que van a ser analizados. A continuación, se especifica el coeficiente de consistencia interna con type=. Con la opción 'alpha' se obtiene α calculado a partir de la Ecuación 4 y el estimador ULS. La opción 'alpha-CFA' hace referencia al coeficiente omega para medidas esencialmente tau-equivalentes calculado a partir de la Ecuación 2 con un estimador ML. El método usado para estimar el error estándar de medida, y por tanto el IC, se especifica con interval.type=. En el ejemplo, la opción 'bsil' utiliza el método *bootstrap* para calcular el error estándar, y una transformación logística para calcular el IC. El número de réplicas *boostrap* se define en B=. Los resultados pueden verse en la columna Fase 3 de la Tabla 2. De manera alternativa, con muestras grandes, puede utilizarse el método delta que es menos costoso desde el punto de vista computacional, especificando interval.type='mll' para una estimación ML con transformación logística, o interval.type='mlrl' para una estimación MLR con transformación logística.

Como conclusión, todas las estimaciones de la fiabilidad obtenidas se consideraron dentro de las normas aceptadas. Ver el texto principal para más detalles.

Caso 2: Análisis de medidas congenéricas

Los resultados de la Fase 1 y Fase 2 incluidos en las Tablas 1 y 2 se obtuvieron utilizando la tabla de datos del Caso 2 (Case2.txt) y reemplazando C1 por C2 en toda la sintaxis original. Dado que el mejor modelo fue el de medidas congenéricas, cambia la estimación de los coeficientes de fiabilidad obtenidos en la Fase 3. A continuación únicamente se comentan las instrucciones que han sufrido cambios.

La instrucción

```
reliability(CFA_C2cong)
```

proporciona los coeficientes α y omega utilizando los mismos métodos de estimación que en el caso anterior. Obsérvese que, aunque para estimar α se requiere un modelo de medidas esencialmente tau-equivalentes o, al menos, cargas factoriales elevadas, en el listado de resultados no aparece ninguna advertencia. Es, pues, responsabilidad del investigador tomar decisiones acerca de los valores que acabará publicando.

Las estimaciones por intervalo pueden obtenerse aplicando la función ci.reliability() a los datos C2, especificando type='alpha' para α y type='omega' para el coeficiente omega. La opción type='omega' aplica la Ecuación 2 a los parámetros estimados con ML bajo el modelo de medidas congenéricas y constituye el principal cambio con respecto al caso anterior. La instrucción completa es la siguiente:

```
ci.reliability(data=C2, type='alpha', interval.type='bsil', B=500)

ci.reliability(data=C2, type='omega', interval.type='bsil', B=500)
```

Concluimos que tanto α como omega resultan adecuados y dan estimaciones de la fiabilidad parecidas tal y como se esperaba debido a las cargas factoriales altas y homogéneas del modelo de medidas congenéricas. Ver el texto principal para más detalles.

Caso 3: Análisis de medidas con errores correlacionados

Los resultados de la Fase 1 y Fase 2 incluidos en las Tablas 1 y 2 se obtuvieron utilizando la tabla de datos "Case3.txt" y reemplazando C1 por C2 en la sintaxis original. Dado que ni el modelo de medidas esencialmente tau-equivalentes, ni el modelo de medidas congenéricas ajustaron a los datos, la Fase 2 se completó comprobando un nuevo modelo más flexible que contempla términos de error correlacionados. En consecuencia, en la Fase 3 cambia la estimación de los coeficientes.

La correlación entre errores se ha modelizado de la siguiente manera:

```
C3err_corr <- 'Factor1 =~ Y1 + Y2 + Y3 + Y4 + Y5 + Y6
Y4 ~~ Y5
```

```
Y4 ~~ Y6
Y5 ~~ Y6'
```

donde ~~ indica la correlación entre los errores de los ítems Y4, Y5 e Y6.

De nuevo, el modelo definido se estimó y ajustó con la instrucción

```
CFA_C3err_corr <- cfa(C3err_corr, C3, std.lv = TRUE)
summary(CFA_C3err_corr, fit.measures=TRUE)
```

La estimación puntual de α y omega se obtuvo con:

```
reliability(CFA_C3err_corr)
```

Como se discute en el texto principal, el valor de α proporciona una estimación incorrecta de la fiabilidad bajo el modelo de errores correlacionados y se incluye en la Tabla 2 solo con fines ilustrativos. El estimador correcto es el que en el listado de resultados aparece con el nombre de *omega2* y que se obtiene a partir de la Ecuación 5. Finalmente, en la Tabla 2 no se ofrece el IC para omega porque la función ci.reliability()no lo proporciona para modelos con errores correlacionados.

Caso 4: Análisis de datos ordinales

Los resultados de la Fase 1 incluidos en la Tabla 1 se obtuvieron utilizando los datos "Case4.txt" y reemplazando C1 por C4 en la sintaxis original. Aunque los datos proceden de una escala tipo Likert de cinco categorías, fueron tratados como datos ordinales debido a los fuertes efectos techo. En consecuencia, en la Fase 1 se obtuvo la matriz de correlaciones policóricas a partir de la siguiente instrucción:

```
polychoric(C4)
```

En caso de que los ítems fuesen dicotómicos, la matriz de correlaciones tetracóricas se podría obtener con la función tetrachoric ().

La especificación del modelo de medida con categorías ordenadas requiere declarar como ordinales las variables mediante la opción ordered=names() en la función cfa(). El modelo de medidas congenéricas categóricas se define de la siguiente manera:

```
C4cong <- 'Factor1 =~ Y1 + Y2 + Y3 + Y4 + Y5 + Y6'
CFA_C4cong <- cfa(C4cong, C4,std.lv = TRUE, ordered=names(C4))
summary(CFA_C4cong, fit.measures=TRUE)
```

De acuerdo con la naturaleza ordinal de los datos, en la Fase 3 los coeficientes de fiabilidad deben cambiar. La instrucción:

```
reliability(CFA_C4cong)
```

calcula los coeficientes de consistencia interna ordinal, de manera que el coeficiente denominado *alpha* en el listado de salida corresponde al alfa ordinal de la Ecuación 7. El valor etiquetado como *omega3* es el coeficiente no lineal basado en SEM de Green y Yang. El resto de valores de omega deben ser evitados porque no son interpretables para datos categóricos. El estimador correcto de la fiabilidad del Caso 4 es el de fiabilidad no lineal basado en SEM, aunque en la Tabla 2 se incluye también, únicamente con fines ilustrativos, el valor del alfa ordinal.

Finalmente, puede obtenerse la estimación por intervalo de la fiabilidad no lineal basada en SEM con la función ci.reliability()y las opciones type='categorical' para definir la naturaleza categórica de los datos e interval.type='bca' para seleccionar el método bootstrap. La sintaxis completa es la siguiente:

```
ci.reliability(data=C4, type='categorical', interval.type='bca')
```