GRUPO EDITORIAL DE PSICOFUNDACIÓN

Psicothema (2023) 35(1) 05-20

Psicothema

www.psicothema.com • ISSN 0214-9915

Colegio Oficial de Psicólogos del Principado de Asturias

Psicothema

Artículo

Coeficiente Alfa: la Resistencia de un Clásico

Eduardo Doval[1], Carme Viladrich[1] y Ariadna Angulo-Brunet[2]

1 Universitat Autònoma de Barcelona.
2 Universitat Oberta de Catalunya.

INFORMACIÓN

Recibido: Julio 27, 2022
Aceptado: Octubre 14, 2022

Palabras clave:
Consistencia interna
Fiabilidad
Alfa de Cronbach
Omega
Software

RESUMEN

Antecedentes: Durante el siglo XX el coeficiente alfa (α) fue ampliamente utilizado en el cálculo de la consistencia interna de las puntuaciones de los test. Después de identificar algunos malos usos, a principios del siglo XXI se difundieron alternativas, especialmente el coeficiente omega (ω). Actualmente α resurge como una opción aceptable. Método: Revisamos aportaciones académicas, hábitos de publicación en revistas y recomendaciones de textos normativos con el fin de identificar buenas prácticas en la estimación de la fiabilidad de consistencia interna. Resultados: Para guiar el análisis, proponemos un diagrama de decisión en tres fases, a saber, descripción de los ítems, ajuste del modelo de medida del test y elección del coeficiente de fiabilidad de las puntuaciones. Para su ejecución proporcionamos recomendaciones sobre el uso de los programas R, Jamovi, JASP, Mplus, SPSS y Stata. Conclusiones: Tanto α como ω son adecuados para ítems que se distribuyen de forma aproximadamente normal y medidas aproximadamente unidimensionales y congenéricas sin cargas factoriales extremas. Cuando los ítems tienen otra distribución, un fuerte componente específico o sus errores están correlacionados, resultan más adecuadas variantes de ω. Algunas de ellas requieren diseños específicos de obtención de datos. A nivel práctico recomendamos un uso crítico del software.

Coefficient Alpha: The Resistance of a Classic

Keywords:
Internal consistency
Reliability
Cronbach's alpha
Omega
Software

ABSTRACT

Background: During the 20th century the alpha coefficient (α) was widely used in the estimation of the internal consistency reliability of test scores. After misuses were identified in the early 21st century alternatives became widespread, especially the omega coefficient (ω). Nowadays, α is re-emerging as an acceptable option for reliability estimation. Method: A review of the recent academic contributions, journal publication habits and recommendations from normative texts was carried out to identify good practices in estimation of internal consistency reliability. Results: To guide the analysis, we propose a three-phase decision diagram, which includes item description, fit of the measurement model for the test, and choice of the reliability coefficient for test score(s). We also provide recommendations on the use of R, Jamovi, JASP, Mplus, SPSS and Stata software to perform the analysis. Conclusions: Both α and ω are suitable for items with approximately normal distributions and approximately unidimensional and congeneric measures without extreme factor loadings. When items show non-normal distributions, strong specific components, or correlated errors, variants of ω are more appropriate. Some require specific data gathering designs. On a practical level we recommend a critical approach when using the software.

Cómo citar: Doval, E., Viladrich, C., & Angulo-Brunet, A. (2023). Coeficiente Alfa: la Resistencia de un Clásico. *Psicothema, 35*(1), 05-20. https://doi.org/10.7334/psicothema2022.321
Autor para dirigir correspondencia: Carme Viladrich, carme.viladrich@uab.cat

En las ciencias sociales y de la salud es habitual usar cuestionarios o test (en adelante nos referiremos a test) para puntuar a las personas en un constructo o variable latente. La puntuación de un test se define, frecuentemente, como la suma o la media de las respuestas de cada persona a los ítems del test y las inferencias sobre el constructo deben basarse en sólidas propiedades psicométricas de dicha puntuación. Entre otras, deben aportarse evidencias de su fiabilidad, tal y como se indica en la norma 2.3 de los *Estándares para Pruebas Educativas y Psicológicas* (American Educational Research Association et al., 2014). Una de las formas de aportar esta evidencia es calcular la fiabilidad de consistencia interna en la que se centra este trabajo.

La fiabilidad de la consistencia interna de la puntuación de un test se basa en el grado de asociación entre las respuestas a sus ítems obtenidas con una única administración del test a un grupo de personas. Su cálculo es muy sencillo en un modelo de medida idealizado, en el que todos los ítems evalúan un único constructo (unidimensionalidad) con la misma capacidad discriminativa (medidas esencialmente tau-equivalentes reflejadas en cargas factoriales iguales). Además, los errores de medida, que se consideran presentes en todas las evaluaciones, son aleatorios y no están relacionados (errores independientes). Por otra parte, las personas evaluadas conforman un grupo con diferencias individuales apreciables en el constructo pero homogéneo respecto a otras características. Este escenario ideal se complementaría con respuestas de una muestra grande, sin datos faltantes, a un test largo compuesto por ítems con un formato homogéneo.

En entornos realistas, los modelos de medida son más complejos. Los ítems del test pueden medir diferentes constructos (multidimensionalidad), o medir el mismo constructo principal, reflejado en un factor general, con algunos de los ítems agrupados en factores secundarios, o bien mostrando errores correlacionados (unidimensionalidad esencial; e.g., debida a ítems redactados de forma similar o bien a ítems que además de la dimensión principal miden otras dimensiones menores). Además, los ítems suelen mostrar una capacidad de discriminación diferente (medidas congenéricas reflejadas en cargas factoriales diferentes) y pueden mostrar una varianza específica no compartida con otros ítems y no asimilable al error de medida (e.g., ítems que evalúan diferentes facetas de un mismo constructo). Además, las personas evaluadas pueden presentar otras características que inducen a la heterogeneidad o incluso pueden pertenecer a clases o grupos bien definidos. Ante cualquiera de estas complejidades, el escenario ideal descrito anteriormente deja de ser realista, por no hablar de los datos perdidos, un test corto o un tamaño de la muestra pequeño.

La mejor manera de acomodar la estimación de la fiabilidad a las condiciones reales ha sido objeto de un intenso debate desde finales del siglo pasado. Para situar el debate, en los siguientes párrafos se exponen los conceptos básicos de tres teorías de la medida que sustentan el cálculo de un coeficiente de consistencia interna.

Según la Teoría Clásica de los Test (TCT; Gulliksen,1950; véase también Muñiz, 2018; Sijtsma y Pfadt, 2021), se espera que las respuestas de las personas a los ítems del test reflejen correctamente las diferencias individuales en el constructo produciendo variabilidad en las puntuaciones. Esta variabilidad entre las personas es el objetivo de la medida y se denomina varianza sistemática o verdadera. Las respuestas de las personas también dependerán de muchos otros factores menores presentes en la evaluación, como la fatiga o la motivación, que producen una variabilidad impredecible denominada error. Si los errores de los ítems son independientes entre sí y respecto a la puntuación verdadera, la varianza total de la suma o promedio de los ítems es la suma de la varianza sistemática más la varianza del error. Los coeficientes de fiabilidad pretenden cuantificar la proporción de varianza sistemática presente en la varianza total y, por tanto, toman valores entre 0 y 1, siendo preferibles los valores altos pero no los valores extremos próximos a 1.

La estimación de la fiabilidad de la consistencia interna más utilizada la proporciona el coeficiente alfa de Cronbach (α), formulado, entre otros autores, por Cronbach (1951) en el marco de la TCT (Gulliksen, 1950) para el escenario ideal representado anteriormente: medidas unidimensionales, esencialmente tau-equivalentes, con errores independientes. Se pueden utilizar varias fórmulas equivalentes para calcular α pero en esencia, se trata de un cociente entre la varianza sistemática y la varianza total. En el numerador el promedio de las covarianzas multiplicado por el número de ítems al cuadrado cuantifica la varianza sistemática. En el denominador, la suma de elementos de la matriz de varianzas-covarianzas de los ítems cuantifica la varianza total. Cuando la puntuación total es la suma (o el promedio) de los ítems y los datos se aproximan al escenario idealizado descrito anteriormente, la relación resultante es una buena estimación de la proporción de varianza sistemática presente en la varianza de la puntuación observada.

El coeficiente α es también un caso particular de los coeficientes de correlación intraclase derivados de la Teoría de la Generalizabilidad (TG; Cronbach et al., 1963, véase también Brennan, 2001). El objetivo principal de la TG es desentrañar las fuentes de error identificables que contribuyen a la varianza del error en la TC y las diferentes medidas de fiabilidad se derivan del ANOVA de medidas repetidas con efectos aleatorios. Entre los coeficientes de correlación intraclase más conocidos, el coeficiente de consistencia para las medidas promedio es igual a α. Este se basa en la covariación entre las medidas repetidas y no debe ser confundido con el coeficiente de concordancia absoluta para las medidas promedio que incluye el requisito adicional de la igualdad de medias entre las medidas repetidas.

Además, la fiabilidad de la consistencia interna también puede derivarse del Análisis Factorial (AF; Thurstone, 1947; véase también Brown, 2015; Ferrando et al., 2022). Si los ítems miden un único factor con errores no correlacionados, las respuestas a los ítems pueden ser explicadas por una parte común (o carga factorial) más una parte única (o unicidad). Asumiendo puntuaciones factoriales estandarizadas, McDonald (1999) definió el coeficiente omega (ω) como la relación entre la varianza común, o cuadrado de la suma de las cargas factoriales, y la varianza total, o cuadrado de la suma de las cargas factoriales más la suma de las unicidades. Este coeficiente también se ha denominado fiabilidad de una puntuación compuesta (Raykov, 1997a), ω_{total} (Revelle y Zinbarg, 2009) y ω_u (Flora, 2020).

Por último, los modelos de medida AF y TCT son equivalentes cuando las cargas factoriales se asimilan con la varianza verdadera y las unicidades se asimilan con la varianza del error (e.g., Green y Yang, 2015). Entonces, α es un caso particular de ω obtenido con datos unidimensionales, errores no correlacionados y cargas factoriales iguales para todos los ítems (medidas esencialmente tau-equivalentes). Por otro lado, los datos unidimensionales con errores no correlacionados y cargas factoriales diferentes para algunos ítems (medidas congenéricas), proporcionarán un valor de más bajo que el de ω. Ambos valores, α y ω, son iguales o inferiores a la fiabilidad

poblacional y, por lo tanto, salvo por variabilidad muestral, los dos son límites inferiores de la fiabilidad. Como ambos coeficientes son estimaciones basadas en muestras, sus valores deben ir acompañados de intervalos de confianza (Oosterwijk, et al., 2019).

A finales del siglo XX, α no tenía una competencia clara como estimador de la fiabilidad, aunque ya se denunciaban algunos usos abusivos (e.g., Cortina, 1993; Schmitt, 1996). A partir de entonces, ha ido creciendo un animado debate sobre cuál es el mejor estimador de la fiabilidad de la consistencia interna. Presentaremos los aspectos de este debate relacionados con la investigación metodológica, los hábitos de publicación de las revistas científicas y la posición de las instituciones normativas.

En la primera década del siglo XXI, se dedicó una gran cantidad de investigación metodológica a identificar y difundir alternativas a α como mejores estimadores del límite inferior de la fiabilidad. Un primer grupo de especialistas abogó por los coeficientes derivados del AF, como ω (Green y Yang, 2009; Raykov, 1997a; Yang y Green, 2011), otros defendieron los coeficientes que no se basan en un modelo de medida específico como el coeficiente glb (por sus siglas en inglés de *greatest lower bound*; Sijtsma, 2009), mientras que un tercer grupo defendía el uso de varios coeficientes para expresar diferentes aspectos de la fiabilidad de consistencia interna (Bentler, 2009; Zinbarg et al., 2005). Más tarde, el trabajo de McNeish (2018) publicado en la revista *Psychological Methods*, defendía el uso de varios coeficientes, pero desaconsejando explícitamente α. Mientras tanto, los estudios de simulación han ido desplazando el debate desde cuál es la mejor estimación del límite inferior hasta cuál es la estimación más precisa de la fiabilidad poblacional. Considerando un gran número de modelos de medida, no se han encontrado diferencias apreciables en la precisión alcanzada por α y ω (e.g., Edwards et al., 2021; Gu et al., 2013; Raykov y Marcoulides, 2015). Además, algunos coeficientes que no se basan en un modelo de medida han mostrado, en estudios de simulación, un comportamiento inaceptable cuando se comprueba su precisión (Edwards et al., 2021; Sijtsma y Pfadt, 2021). Por último, el coeficiente ω ha sido criticado por el gran número de decisiones intermedias necesarias para obtenerlo (Davenport et al., 2016) y por el hecho de que ω no se refiere a un único indicador sino a toda una familia de coeficientes, lo que puede dificultar la comparación entre estudios (Scherer y Teo, 2020; Viladrich et al., 2017). Basándose en algunas de estas razones, cada vez son más las voces que reclaman la vuelta a α, incluso procedentes de autores que anteriormente habían defendido otras alternativas (Raykov et al., 2022; Sijtsma y Pfadt, 2021). Otras posiciones consideran que el uso de α y de otros coeficientes debe depender del cumplimiento de sus supuestos subyacentes (Green y Yang, 2015; Raykov y Marcoulides, 2016; Savalei y Reise, 2019; Viladrich et al., 2017). Además, se ha propuesto como buena práctica la publicación conjunta de α y otros coeficientes alternativos (e.g., Bentler, 2021; Revelle y Condon, 2019).

Este debate entre los académicos se ha reflejado de forma ambigua en los hábitos de publicación de las revistas científicas. Flake et al. (2017) encontraron que el 73% de los 301 artículos que analizaron publicaron α. La encuesta realizada por Hoekstra et al. (2019) a 664 investigadores que publicaron α en revistas relevantes de diferentes campos, proporcionó algunas explicaciones a este hecho. Aunque el 88% declaró conocer alternativas a α, el 74% dijo que publicaba α porque esa era la práctica habitual en su campo, el 53% sigue publicando α porque cree que así se lo exigirá la revista o el proceso de revisión, y el 43% dijeron que ese era el coeficiente que les enseñaron a calcular durante su formación científica.

Buscando posiciones extremas en los hábitos de publicación, hemos revisado los trabajos científicos que citan el trabajo de McNeish (2018). Debido a su posición contraria al uso esperábamos que estos trabajos publicaran principalmente otros coeficientes. En el momento de escribir nuestro texto (septiembre de 2022) encontramos 696 trabajos citados en la Web of Science de los que pudimos consultar 672. Entre los 598 que publicaron datos empíricos, 79 (13,2%) informaron solo de α; 207 (34,6%) publicaron α y otro coeficiente, en general ω; 251 (42,0%) solo publicaron ω; 21 (3,5%) publicaron un coeficiente ω y otro distinto de α; 28 (4,7%) de un coeficiente distinto de ω o α; y 12 (2,0%) no informaron de ningún coeficiente de fiabilidad. Es decir, prácticamente la mitad de los trabajos que citan el artículo de McNeish publicaron α a pesar del consejo contrario de este autor.

Desde posiciones normativas, el manual de publicación de la *American Psychological Association*, recomendaba indirectamente publicar α hasta su versión 6. En la versión 7 (American Psychological Association, 2020) se promueve explícitamente informar sobre la validación del modelo de medida previo al cálculo del coeficiente de fiabilidad, siguiendo las recomendaciones de Appelbaum et al. (2018) y Slaney et al. (2009). Además, la posibilidad de publicar α u otros coeficientes como ω, conjuntamente o por separado, ha sido adoptada por las comisiones europeas de evaluación de test (e.g., el cuestionario CET-R [Cuestionario de Evaluación de Tests Revisado de la Comisión de Test del Colegio Oficial de la Psicología de España], Hernández et al., 2016; el modelo EFPA, [Federación Europea de Asociaciones de Psicólogos], Evers et al., 2013; el modelo COTAN [Comité Holandés de Tests], Evers et al., 2015). Asimismo, en los métodos para revisiones sistemáticas y meta-análisis, que pueden considerarse normativos para los estudios primarios, se desaconseja el uso acrítico de α. En cambio, se promueve el análisis previo del modelo de medida (Prinsen et al., 2018) y la publicación del coeficiente de fiabilidad más adecuado a las características de los datos (Sánchez-Meca et al., 2021).

Así, el debate parece haberse decantado a favor de α y ω sobre los coeficientes no basados en modelos de medida. Una de las principales ventajas de ω sería su adaptabilidad a primera vista a modelos de medida más sofisticados derivados del AF, mientras que la mayor ventaja de α sería su simplicidad. Sin embargo, tanto si se utiliza α como otro coeficiente, la elección de esos coeficientes nunca debe ser irreflexiva y siempre debe poder justificarse.

A la luz de lo que parece ser una nueva oportunidad para α, nos propusimos revisar nuestras directrices para el uso del coeficiente de fiabilidad más adecuado en diferentes escenarios analíticos (Viladrich et al., 2017). En aquel trabajo distinguíamos el coeficiente a utilizar en función de la naturaleza de los datos y del modelo de medida. Mantenemos nuestra alineación con la opinión de que la elección de un coeficiente de fiabilidad depende del modelo de medida que mejor se ajuste a los datos (véase también Green y Yang, 2015; Raykov y Marcoulides, 2016; Savalei y Reise, 2019). Por el contrario, nuestra posición difiere de la de quienes han defendido recientemente la idea de que ω debería sustituir de forma rutinaria a α como indicador de la fiabilidad de consistencia interna (e.g., Flora, 2020; Goodboy y Martin, 2020; Komperda et al., 2018). Nuestro trabajo también difiere de aquellos que sugieren que el coeficiente adecuado puede ser obtenido de manera rutinaria utilizando un software de "señalar y clicar" (e.g., Kalkbrenner, 2021).

Así, el primer objetivo de este trabajo es revisar los criterios para la toma de decisiones a la hora de estimar la fiabilidad de consistencia interna. Para ello examinaremos la investigación que muestra cuándo α es adecuado y cuáles son sus mejores alternativas,

especialmente ω, cuando α no es adecuado. En nuestra opinión, la elección del coeficiente de fiabilidad es el resultado de varias decisiones sucesivas que resumimos en un diagrama de flujo estructurado en tres fases analíticas. El segundo objetivo de este trabajo es facilitar la aplicación de estos criterios cuando se utiliza un software estadístico para la estimación de la fiabilidad. Para ello, compararemos algunos de los programas informáticos más comunes o fáciles de usar para llevar a cabo las tres fases analíticas propuestas que terminan con una estimación adecuada de los coeficientes de fiabilidad. Por último, se derivarán algunas conclusiones y recomendaciones de importancia para las personas que analizan datos y que revisan artículos científicos.

Cuándo y por qué Utilizar α y/u ω

Los usos recomendados de α y de ω se basan en varios argumentos. Estos argumentos tienen que ver con la naturaleza de las variables o de los grupos de personas, con las cargas factoriales, con la dimensionalidad del test o con los modelos de medida adecuados para describir las respuestas. A continuación, se revisa a la vista de la literatura más reciente si cada uno de estos aspectos condiciona, o no, el uso de α u ω.

Continuidad y normalidad

En un escenario ideal α puede utilizarse para estimar la fiabilidad de la suma o el promedio de las respuestas en una escala continua. Considerando que la continuidad no es un requisito, Chalmers (2018) propone utilizar α también si las escalas de respuesta son ordinales politómicas o incluso dicotómicas. En cambio, los resultados de Xiao y Hau (2022) muestran sesgos que pueden ser elevados en este caso. Además, hay que tener en cuenta que, con las escalas de respuesta ordinales, el modelo de medida suele construirse invocando variables latentes continuas para las que se han observado respuestas discretas (Zumbo y Kroc, 2019). Por esta razón, cuando se enfrentan a formatos de respuesta ordinal, algunos autores optan por utilizar las versiones ordinales de α u ω en las que la fiabilidad se calcula en la métrica de las variables latentes continuas (Elosua y Zumbo, 2008; Gadermann et al., 2012; Zumbo et al., 2007) y otros optan por la versión no lineal o categórica de ω ($\omega_{categórico}$) en la que el coeficiente de fiabilidad se calcula en la métrica de las variables observadas discretas (Green y Yang, 2009). Debido a su métrica, y como argumentamos en un trabajo anterior (Viladrich et al., 2017), cuando la escala de respuesta es ordinal, en el presente trabajo nos inclinamos por utilizar el coeficiente α de Cronbach o el coeficiente $\omega_{categórico}$.

Además, en principio, la forma de la distribución de las respuestas de los ítems del test y, en particular, la normalidad, no es un supuesto necesario para el uso de α (Raykov y Marcoulides, 2019). Sin embargo, se sabe que la distribución de los ítems puede afectar a la estimación de las covarianzas y correlaciones entre ellos y, por tanto, a la estimación de α. Mientras que en presencia de una curtosis positiva subestima la fiabilidad, en presencia de curtosis negativa puede sobrestimarla ligeramente, sesgos que se atenúan en muestras grandes de, por ejemplo, 1000 casos (Olvera et al., 2020). Además, en presencia de asimetría la estimación de la fiabilidad también está sesgada hacia valores bajos (Kim et al., 2020). Peor aún, si la desviación de la normalidad es notable, como ocurre con los efectos techo o suelo, todos los resultados relacionados con la fiabilidad de la consistencia interna se ven afectados, desde la matriz de correlaciones policóricas (Foldnes y Grønneberg, 2020) y la determinación de la dimensionalidad del test (Christensen et al., 2022) hasta el cálculo de ω (Yang y Xia, 2019), para tratar estos casos se han desarrollado coeficientes específicos (Foster, 2021) pero, como el propio autor reconoce, su uso es limitado porque se basan en supuestos muy exigentes sobre el modelo de medida, y su eficacia en comparación con α y ω todavía no ha sido suficientemente investigada. En este momento, sería más seguro optar por un modelo de medida no lineal basado en la Teoría de Respuesta al Ítem (TRI) y derivar de él un coeficiente de fiabilidad asimilable a los de consistencia interna derivados de la TCT (e.g., Culpepper, 2013; Kim y Feldt, 2010; Raykov et al., 2010). Por otra parte, si las irregularidades en la distribución se deben a una baja selección de algunas categorías de respuesta, se puede recurrir a la solución clásica de agrupar las categorías antes de iniciar el análisis de fiabilidad (e.g., Agresti, 1996; DiStefano et al., 2020).

Grupos homogéneos

En cuanto a la homogeneidad de las personas que contestan al test, cuando las poblaciones se estructuran en clases heterogéneas, las estimaciones de los parámetros podrían estar sesgadas y sus errores estándar ser incorrectos, por lo que se recomienda identificar las clases presentes y calcular la fiabilidad por separado en cada una de ellas (Raykov et al., 2019). Si la heterogeneidad se debe a una estructura multinivel, Lai (2021) propone utilizar modificaciones de α y ω, aunque su comportamiento en datos reales no ha sido suficientemente estudiado.

Cargas factoriales homogéneas

La diversidad en las cargas factoriales es la principal fuente de diferencia entre α y ω. Unas cargas factoriales no homogéneas podrían derivar del contenido de los ítems o simplemente de diferencias considerables entre las varianzas de los ítems (Graham, 2006). Si el modelo de medida es unifactorial sin errores correlacionados, el coeficiente α subestima la fiabilidad incluso si solo hay una carga factorial que es muy diferente de las demás (Raykov, 1997b), especialmente cuando el número de ítems es pequeño. Sin embargo, cuando las cargas factoriales son, en promedio, de .70 y las discrepancias de las cargas factoriales entre ellas son, en valor absoluto, inferiores a .20, las diferencias entre α y ω son mínimas (Raykov y Marcoulides, 2015). Los recientes estudios de simulación publicados por Edwards et al. (2021) sugieren que incluso con discrepancias más extremas, por ejemplo, con cargas factoriales entre .20 y .80 en muestras de 100 o más casos, las subestimaciones que se producen son pequeñas, con una media de .02 para 12 ítems y de .04 para 6 ítems. Incluso se puede observar una posición más radical en Raykov et al. (2022).

Además, las diferencias entre los dos coeficientes se reflejan generalmente a partir del tercer decimal si los datos proceden de una fiabilidad poblacional razonable desde un punto de vista práctico (próxima a .80). Solo se obtienen diferencias más grandes cuando la fiabilidad poblacional es extremadamente baja. Además, Deng y Chan (2017), y Hussey y Hughes (2020), analizando datos

reales, informan de diferencias entre α y ω en el tercer decimal. Es decir, a pesar de las llamadas de atención relacionadas con las cargas factoriales diferentes, en la mayoría de los casos no habría ninguna diferencia práctica entre utilizar un coeficiente u otro.

Test multidimensionales

Por último, los coeficientes α y ω no son adecuados para los test multidimensionales que miden diferentes constructos que no comparten un factor general. No obstante, una vez identificados los diferentes factores, estos coeficientes pueden calcularse para cada subescala por separado (Bentler, 2021; Flora, 2020; Prinsen et al., 2018; Sijtsma y Pfadt, 2021).

Modelo de medida

Las principales preocupaciones para el uso de α y ω surgen cuando (a) la varianza única de algunos ítems no es asimilable al error de medida, (b) algunos errores de los ítems están correlacionados, o (c) se identifican factores menores además del factor general. Dicho de otro modo, cuando los resultados del AF vulneran los supuestos de la TCT y los dos modelos dejan de ser equivalentes, no solo α, sino también α, están en duda.

El primero de los escenarios planteados es común a los test diseñados para medir conceptos amplios con pocos ítems, como por ejemplo los tests breves de personalidad. La especificidad del contenido de los ítems, que se reflejará en la varianza no compartida o unicidad, es necesaria para lograr la medida del constructo y, por tanto, no puede asimilarse al error de medida. En este caso, tanto α como ω subestimarán notablemente la fiabilidad del test hasta el punto de invalidar la conclusión clásica de que la fiabilidad es el límite superior de la validez (McCrae, 2015).

Más preocupante aún es la presencia de errores correlacionados entre ítems o de factores menores. Son frecuentes en los test y pueden deberse a similitudes en el significado de algunos ítems, efectos de orden, efectos de formato de respuesta (Bandalos, 2021; Weijters et al., 2009) o la influencia de factores específicos identificables más allá del factor común (Rodriguez et al., 2016a, 2016b). En estos escenarios, los coeficientes α y ω algunas veces pueden subestimar y otras sobreestimar el valor poblacional de la fiabilidad, perdiendo así la tan apreciada garantía de ser estimaciones conservadoras de la fiabilidad (Bentler, 2021; Raykov, 2001).

Se han propuesto tres tipos de soluciones para hacer frente a estos problemas. Todas ellas requieren el juicio del investigador. Una opción es considerar que la fiabilidad del constructo se refiere únicamente a la varianza común entre todos los ítems. Los demás componentes de la varianza, incluyendo la varianza específica y las covarianzas residuales o los factores de grupo, se considerarán parte del error de medida y, en consecuencia, el cálculo de los coeficientes de consistencia interna se corregirá (reducirá) incluyéndolos solo en el denominador. Si se adopta este curso de acción, fórmulas como $\omega_{jerárquico}$ (Zinbarg et al., 2005) u ω corregida por errores correlacionados (Raykov, 2004) serán útiles. Hay que tener en cuenta que las cargas factoriales para los cálculos deben derivarse de un AF con el modelo de medida adecuado, por ejemplo, el modelo bifactorial con un factor común y algunos factores de grupo, o bien el modelo de un factor con algunos errores correlacionados.

Otra opción es considerar toda la varianza compartida como varianza verdadera, incluyendo la varianza común y las covarianzas entre algunos ítems o los factores de grupo. En consecuencia, ambas se incluirán en el numerador y el denominador del coeficiente de fiabilidad de consistencia interna. Si se acepta esta opción, la fórmula para estimar la fiabilidad de consistencia interna sería ω con las cargas factoriales comunes y de grupo obtenidas de un modelo bifactorial (Revelle y Zinbarg, 2009; Zinbarg et al., 2005) o también α.

Una tercera opción es diferenciar estos componentes de la varianza midiendo los predictores de la varianza y/o las covarianzas residuales. La varianza predicha por estas covariables se convertirá en varianza explicada separada del error de medida aleatorio. Esto puede lograrse de varias maneras. En diseños transversales Bentler (2017) propone la incorporación de variables auxiliares. Un enfoque preferible en diseños longitudinales sería utilizar conceptos de series temporales como los errores autocorrelacionados (Green y Hershberger, 2000) o la identificación de factores específicos de los ítems basados en medidas repetidas (Raykov, 2007). En cualquier caso, si el investigador quiere diferenciar los componentes de la varianza, la decisión principal ya no se reduce a la elección de la mejor fórmula para la estimación de la fiabilidad, sino que abarca el diseño de recogida de datos registrando las variables auxiliares en un diseño transversal o las medidas repetidas en un diseño longitudinal. El coeficiente de fiabilidad se calculará a partir de un coeficiente ω corregido (aumentado) incluyendo la parte específica predicha como varianza verdadera tanto en el numerador como en el denominador (ω + u ω con corrección de la especificidad ω según Bentler, 2017; ω_i según Raykov, 2007) o incluyendo la variabilidad atribuible a los errores autocorrelacionados solo en el denominador (Green y Hershberger, 2000).

Como resumen de este apartado, la Tabla 1 muestra, para un conjunto de escenarios determinados por la definición de varianza verdadera (filas), los coeficientes de fiabilidad α u ω que recomendamos en función de si los datos pueden tratarse como cuantitativos (columna 3) o como ordinales (columna 4). Las recomendaciones son aplicables a ítems de formato homogéneo con respuestas en una escala de categorías ordenadas, dicotómicas o politómicas, para la estimación de la fiabilidad de la suma o promedio de las respuestas observadas de los ítems, no de las hipotéticas respuestas continuas subyacentes ni de las puntuaciones factoriales.

El primer escenario es uno de los más comunes: el análisis de ítems que miden un único factor aunque sus cargas factoriales no sean especialmente homogéneas. En esta situación, el uso de α u ω, en su versión cuantitativa u categórica según el tipo de ítems analizados, estaría perfectamente justificado, dando lugar a valores muy similares. En este escenario, si algunas cargas factoriales son extremas, el curso de acción a tomar es todavía objeto de discusión (Edwards et al., 2021, Raykov et al., 2022). Los resultados del estudio de simulación de Edwards et al. (2021) con datos cuantitativos desaconsejarían el uso de α con cargas factoriales fuera del intervalo 0.2 – 0.8, aunque por el momento no hemos encontrado estudios similares para el caso de datos ordinales. Creemos que la propuesta más conservadora en este caso sería el uso de ω.

Tabla 1.
Uso Recomendado de los Coeficientes α y ω para Obtener la Fiabilidad de Consistencia Interna en Diferentes Escenarios.

Escenario	Varianza verdadera	Coeficiente recomendado para la fiabilidad de la suma o promedio de los ítems	
		Cuantitativos: Cinco o más categorías de respuesta y relación lineal con errores normales	Ordinales: Cuatro o menos categorías de respuesta y relación linealizable
1	modelo unidimensional (varianza verdadera = varianza común)	• α (Cronbach, 1951)** • ω (McDonald, 1999)	• α (Cronbach, 1951)** • $\omega_{categórico}$ (Green y Yang, 2009)
2	modelo esencialmente unidimensional (varianza verdadera = varianza común + varianza de factores menores)	• α (Cronbach, 1951) • ω_{total} derivado de un modelo bifactorial (Zinbarg et al., 2005)	• α (Cronbach, 1951) • Versión categórica del coeficiente ω pendiente de desarrollar
3	modelo esencialmente unidimensional (varianza verdadera = varianza común)	• $\omega_{jerárquico}$ (Zinbarg et al., 2005) • $\omega_{corregido}$ por errores correlacionados (Raykov, 2004)	• $\omega_{h\text{-}cat}$ (Flora, 2020)
4	modelo unidimensional (varianza verdadera = varianza común + varianza específica)	• ω_i (Raykov, 2007)	• Versión categórica del coeficiente ω pendiente de desarrollar

Nota: α = alfa de Cronbach; ω = omega (también u, total o fiabilidad de la puntuación compuesta); ω_i = omega con corrección por especificidad (también ω + o specificity enhanced ω; $\omega_{categórico}$ (también ωno lineal). $\omega_{h\text{-}cat}$ = omega jerárquico para datos categóricos. ** En el escenario 1, estudios de simulación con datos cuantitativos favorecen el uso de omega cuando las cargas factoriales son extremas. Faltan estudios de simulación equivalentes para datos categóricos.

El resto de los escenarios proporcionan soluciones para otras formas de concebir la varianza verdadera. En el segundo escenario, en el que se encuentran algunos factores menores definidos por correlaciones entre ítems no explicadas por el factor general (unidimensionalidad esencial) y esta variabilidad se considera como varianza verdadera, ω debe derivarse de un modelo bifactorial y considerar la varianza común y de los factores menores como varianza verdadera. Tal como se refleja en la Tabla 1, este coeficiente ω está desarrollado para datos cuantitativos, pero hasta donde conocemos, todavía no se ha desarrollado una versión del mismo para datos categóricos u ordinales. En el tercer escenario tratamos la otra opción frente a la multidimensionalidad esencial, en la que la variabilidad de los factores menores se considera varianza de error. En este caso, consideramos más apropiado utilizar $\omega_{jerárquico}$ u ω corregido por errores correlacionados para datos cuantitativos y $\omega_{h\text{-}cat}$ para datos categóricos u ordinales. El cuarto escenario se refiere a los ítems cuya especificidad se considera como varianza verdadera dentro de un modelo unidimensional. En este caso pensamos que, en datos cuantitativos, ω_i es la opción correcta para estimar la fiabilidad, aunque solo será posible si se ha previsto en el diseño de la recogida de datos. De nuevo, y hasta donde conocemos, este tipo de coeficiente no ha sido desarrollado para datos categóricos u ordinales.

La Elección de un Coeficiente de Fiabilidad: Un Análisis en Tres Fases

De lo elaborado hasta aquí debería haber quedado claro que desaconsejamos el análisis de la fiabilidad de consistencia interna de un test eligiendo la instrucción por defecto en el software preferido. Por el contrario, compartimos con otros trabajos la idea de que este análisis implica un proceso complejo pero necesario (Liddell y Kruschke, 2018; Savalei y Reise, 2019). Estructuramos este proceso en tres fases en las que se toman decisiones sucesivamente: (1) la descripción estadística de los ítems; (2) el ajuste del modelo de medida para el test y (3) la estimación de la fiabilidad de consistencia interna de la(s) puntuación(es) del test. Este trabajo se centra en la tercera fase, pero, como hemos visto, la elección razonada del coeficiente de fiabilidad en esta tercera fase depende de las decisiones tomadas en las dos primeras. Por lo tanto, a continuación, se ofrecen también algunas pautas para abordar las dos primeras fases. Las tres fases propuestas se representan en la Figura 1. La trayectoria vertical resaltada y sombreada en la parte izquierda de la Figura 1 representa el análisis que conduce al cálculo del coeficiente α tal y como se recomienda en la Tabla 1 para el primer escenario que es el más común. Las alternativas más complejas analizadas en este documento se representan en color más claro y sin sombreado.

Fase 1: Descripción Estadística de los Ítems

El objetivo de esta primera fase es conocer la distribución de las respuestas a los ítems, detectar la posible presencia de datos perdidos e inspeccionar los subgrupos de personas e ítems en busca de posibles patrones que puedan orientar la modelización que se llevará a cabo en la siguiente fase.

Fase 1a: Completitud de los Datos

La descripción univariante de los ítems proporciona información sobre la distribución de las respuestas, incluyendo los posibles valores perdidos. Si los datos están completos, se puede continuar con el análisis según lo previsto. Si se detectan algunos valores perdidos, se recomienda utilizar la imputación múltiple tanto si los datos que se analizan como cuantitativos (Ferrando et al., 2022) o como ordinales (Shi et al., 2020). Otras posibilidades son utilizar la estimación de máxima verosimilitud con información completa (FIML) durante la Fase 2 o perfeccionar el análisis según las recomendaciones de los textos especializados (e.g., Enders, 2010). Todas ellas son mejores opciones que eliminar del análisis los casos con datos perdidos (listwise) o tratar los datos perdidos con base en la información bivariada (pairwise), que es lo que se hace por defecto en algunos programas informáticos. Otra cosa muy distinta es que se observen categorías con escasas respuestas o sin ninguna respuesta. No hay forma de inferir este tipo de respuestas no observadas y eso puede suponer un problema para el análisis posterior. Para seguir analizando estos datos de forma categórica u ordinal, se puede optar por colapsar algunas categorías cercanas (e.g., Agresti, 1996; DiStefano et al., 2020). En las fases previas de la investigación, si en la población la probabilidad de elección de algunas categorías de respuesta es muy baja, se puede considerar la posibilidad de reunir una muestra muy grande de personas evaluadas o también de rediseñar la escala de respuesta.

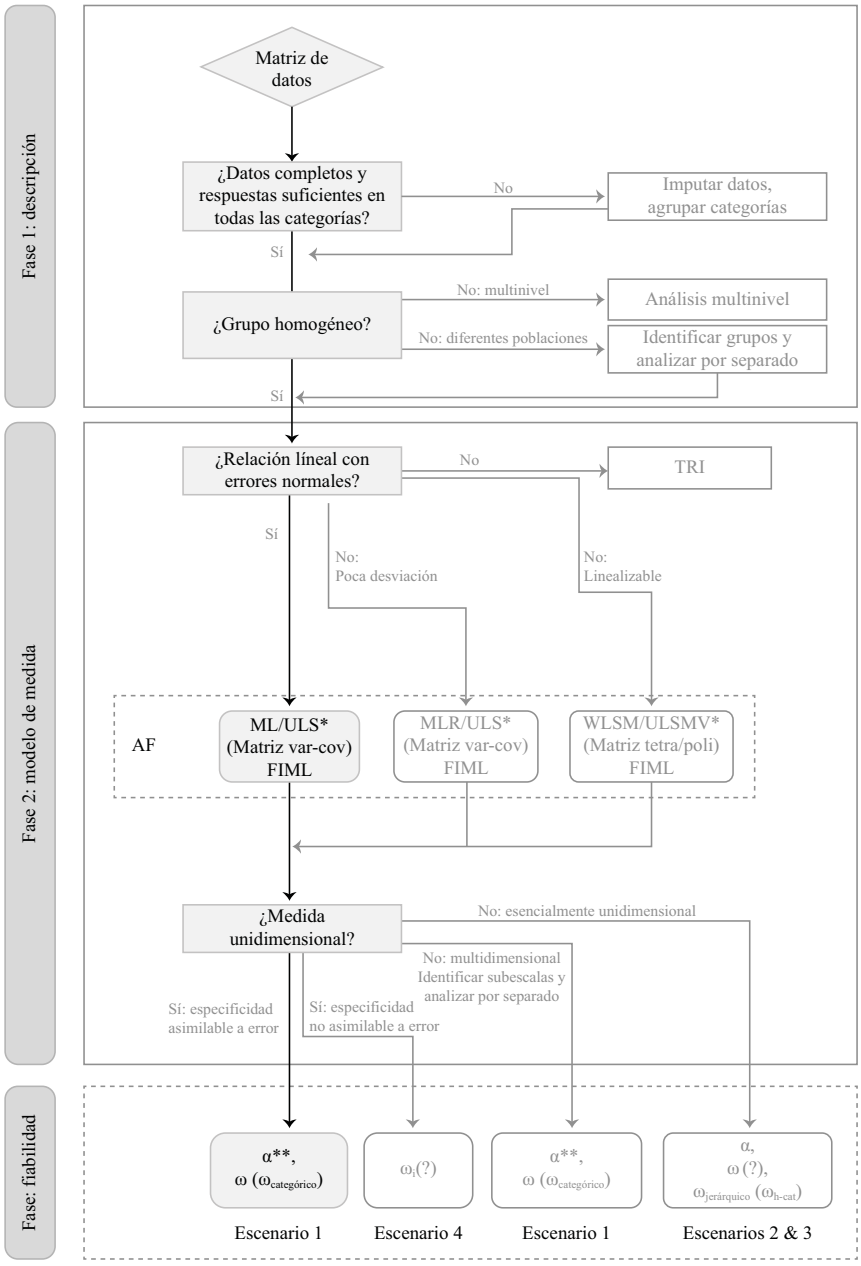

Figura 1.
Diagrama de Toma de Decisiones Para el Coeficiente de Fiabilidad.
Nota: TRI = Teoría de Respuesta al Ítem; AF = Análisis factorial; ML = Estimador de máxima verosimilitud; ULS = Estimador de mínimos cuadrados no ponderados; * = utilizar en muestras pequeñas, var-cov = matriz de varianzas-covarianzas; FIML = Estimador de máxima verosimilitud con información completa; MLR = Estimador de máxima verosimilitud robusto; WLSMV = estimador robusto de mínimos cuadrados ponderados con un estadístico χ^2 ajustado para la media y la varianza; ULSMV = Estimador de mínimos cuadrados no ponderados con un estadístico $\chi 2$ ajustado para la media y la varianza; tetra/pol = matriz de correlaciones tetracóricas o policóricas; Los escenarios se definen en la Tabla 1. α = alfa de Cronbach; ω = omega (también ω_u, ω_{total} o fiabilidad de la puntuación compuesta); ω_i = omega con corrección de la especificidad (también ω + o specificity enhanced ω; $\omega_{categórico}$ (también $\omega_{no\ lineal}$). $\omega_{h\text{-}cat}$ = omega jerárquico para datos categóricos. Los coeficientes que están separados por una coma se pueden presentar juntos o se puede elegir uno de ellos de manera razonada. Entre paréntesis los coeficientes para datos ordinales. ? = coeficiente por desarrollar.

Fase 1b: Homogeneidad de Personas e Ítems

Otra tarea será evaluar si los datos provienen de una población homogénea. Si es así, podemos proceder al análisis según lo previsto. Por otro lado, si el diseño de recogida de datos ha sido multinivel, es aconsejable tratar la heterogeneidad mediante técnicas de análisis multinivel (Cho et al., 2019; Hox et al., 2018). Si la heterogeneidad proviene de datos procedentes de poblaciones diferentes, una opción es continuar el análisis para cada grupo por separado. Si no se conocen las poblaciones subyacentes, pueden identificarse mediante un análisis de conglomerados o incluso mediante un análisis de clases latentes como proponen Raykov et al. (2019).

También resulta útil inspeccionar la homogeneidad de las relaciones entre los ítems. Las relaciones heterogéneas anticipan

posibles desviaciones de la unidimensionalidad que aflorarán en el análisis formal durante la Fase 2. Para los datos cuantitativos, se puede examinar la matriz de varianzas-covarianzas (o la matriz de correlaciones de Pearson). Para los datos categóricos u ordinales sería mejor opción la matriz de correlaciones tetracóricas (dos categorías de respuesta) o policóricas (más de dos categorías de respuesta). La inspección visual de estas matrices puede ser suficiente si el número de ítems no es elevado. De forma más general, la inspección puede realizarse mediante técnicas estadísticas multivariantes como el análisis factorial exploratorio (AFE; e.g., Lloret-Segura et al., 2014), el análisis psicométrico de redes (e.g., Golino y Epskamp, 2017; véase una aplicación práctica en Pons et al., 2017), o el análisis de correspondencias múltiples (e.g., Greenacre, 2017).

El resultado de la Fase 1 es una base de datos para cada población en la que se estudiará formalmente el modelo de medida del test durante la Fase 2.

Fase 2: Análisis del Modelo de Medida del Test

Los objetivos principales de esta fase son determinar la dimensionalidad del test, ya que α y ω solo son adecuados para medidas unidimensionales, y estimar los parámetros que intervienen en el cálculo de ω.

Fase 2a: Especificación del Modelo de Medida

El primer paso será especificar una relación razonable entre los ítems y los factores o variables latentes. Si se supone que las relaciones son lineales y los residuos se distribuyen normalmente, se pueden utilizar técnicas de estimación de información limitada de manera que se simplifica el análisis de la Fase 2b. En cambio, si las relaciones se especifican como no lineales, serán apropiadas las técnicas de estimación de información completa, las mismas técnicas mencionadas anteriormente para tratar los datos faltantes.

Si las categorías de respuesta son cinco o más es razonable tratar los ítems como cuantitativos y linealmente relacionados con las variables latentes siempre que las respuestas a los ítems sigan una distribución normal (Rhemtulla et al., 2012). De hecho, pueden tratarse como normales si los valores absolutos de la asimetría y la curtosis no son mayores que 1 (e.g., Ferrando et al., 2022; Lloret-Segura et al., 2014). Cuando se encuentren desviaciones moderadas de la normalidad, bastarán pequeñas correcciones, que se examinarán en la Fase 2b. En caso contrario, si se detectan desviaciones extremas, como las causadas por los efectos suelo o techo, habrá que considerar un cambio radical de estrategia. En este caso, será aconsejable invocar otras distribuciones de los residuales, como el modelo de Poisson (e.g., Foster, 2020; Muthén et al., 2016, cap. 7) o recurrir a modelos no lineales como los que se describen en el párrafo siguiente.

En cambio, si los ítems se responden en una escala de respuesta de cuatro categorías o menos, ya no es razonable una relación lineal con las variables latentes y, por tanto, es preferible tratar los datos como categóricos u ordinales (Rhemtulla et al., 2012). La relación puede adoptar varias formas, pero en el caso habitual de elegir un modelo de dos parámetros (dificultad del ítem y discriminación del ítem) o el de respuesta graduada (dificultad de las categorías y discriminación del ítem), entonces las relaciones son linealizables calculando coeficientes de correlación policórica o tetracórica. En caso contrario, si el interés es por modelos más complejos, por ejemplo, con más parámetros, la alternativa son los modelos TRI (Culpepper, 2013; Kim y Feldt, 2010).

Una buena práctica es especificar todos los modelos de medida compatibles con la teoría subyacente al constructo, analizarlos uno tras otro y elegir el que mejor se ajuste a los datos y a los fines para los que se va a utilizar el test. Cuando el test pretende medir varios constructos o factores, una secuencia típica de modelos anidados a comprobar es (1) un modelo flexible que permita cargas cruzadas de ítems entre factores, y (2) un modelo restringido en el que los factores sean medidas congenéricas sin cargas cruzadas. Si el test mide solo un constructo, la secuencia se reduce al paso (2) y quizás a la comprobación (3) del modelo de medidas esencialmente tau-equivalentes. Por otro lado, si se sospecha que hay heterogeneidad en las relaciones entre los ítems de un constructo, una secuencia razonable de modelos a comprobar sería (1) un modelo bifactorial, (2) el modelo de medidas congenéricas y, quizás, (3) el modelo de medidas esencialmente tau-equivalentes.

Fase 2b: Estimación de los Parámetros y Ajuste del Modelo

Para la estimación de los parámetros, se puede utilizar el AF de los ítems o los modelos TRI no lineales, siempre que se disponga de datos de muestras amplias. Muchos de los casos pueden resolverse mediante AF utilizando técnicas confirmatorias (AFC) o exploratorias (AFE; e.g., Bovaird y Koziol, 2012). En el caso más sencillo, si los datos son cuantitativos con respuestas de ítems distribuidas normalmente, se recomienda el uso del estimador de máxima verosimilitud (ML por sus siglas en inglés). Como alternativa para desviaciones leves de la normalidad, es preferible el uso del estimador robusto de máxima verosimilitud (MLR). Con datos ordinales y un modelo de dos parámetros o bien de respuesta graduada, el estimador robusto de mínimos cuadrados ponderados con un estadístico χ^2 ajustado para la media y la varianza (WLSMV) se considera una opción adecuada. Siempre puede elegirse la solución general de estimar los parámetros mediante máxima verosimilitud con información completa (FIML) si se acepta un mayor coste computacional.

Si el tamaño de la muestra es pequeño en relación con el número de ítems, una opción preferible para el AF con datos cuantitativos puede ser el estimador de mínimos cuadrados no ponderados (ULS; Ferrando et al., 2022) o bien, para datos ordinales, el de mínimos cuadrados no ponderados con un estadístico χ^2 ajustado para la media y la varianza (ULSMV; Savalei y Rhemtulla, 2013).

El número de casos que se considera un tamaño de muestra pequeño es un tema difícil, pero a modo de guía, los analizados en la literatura son del orden de 100 a 200 casos (e.g., Forero et al., 2009; Savalei y Rhemtulla, 2013).

El resultado de la Fase 2 es el modelo de medida del test que (1) es teóricamente sólido, (2) muestra un buen ajuste a los datos y (3) muestra un mejor ajuste que los modelos alternativos compatibles con la teoría. Con frecuencia, el resultado será un modelo unidimensional, un modelo esencialmente unidimensional o un modelo multidimensional.

Fase 3: Estimación de la Fiabilidad de la Consistencia Interna de la(s) Puntuación(es)

Como se ha visto en los apartados anteriores, la fiabilidad de la consistencia interna de la puntuación de un test con estructura unidimensional y especificidad asimilable al error de medida puede ser estimada tanto con α como con ω que proporcionará valores similares. También se puede optar por informar de ambos tipos de coeficientes. Por el contrario, si la especificidad se considera como varianza verdadera, el coeficiente ω_i reflejará mejor la fiabilidad de la puntuación del test.

Por otro lado, si el modelo de medida es multidimensional, se puede calcular α y/u ω para cada factor por separado. Cuando el modelo es esencialmente unidimensional, nuestra recomendación sería aclarar si se considera toda la parte no común como error de medida, lo que sería más coherente con la publicación del coeficiente $\omega_{jerárquico}$, o si se considera los factores menores como varianza verdadera, lo que sería más coherente con la publicación de ω o incluso α. En este caso también puede ser útil informar sobre ambos tipos de coeficientes (e.g., Green y Yang, 2015).

Por último, en todos los casos, una buena práctica es publicar el intervalo de confianza de los coeficientes de consistencia interna elegidos o la estimación bayesiana de estos coeficientes (Pfadt et al., 2022). En caso de optar por coeficientes alternativos que superan los objetivos de este trabajo aconsejamos consultar literatura especializada. Sería el caso, por ejemplo, de los coeficientes derivados de la TRI, del análisis multinivel, o de otros muchos disponibles (Cho, 2022).

Programas Informáticos Para la Estimación de la Fiabilidad de Consistencia Interna

En esta sección, presentamos las posibilidades actuales de algunos de los programas informáticos más utilizados para llevar a cabo el análisis en tres fases antes descrito. En la mayoría de los casos, el análisis puede desarrollarse por completo utilizando uno o, a lo sumo, dos de ellos. Presentamos el software de código abierto R, Jamovi y JASP, y el software comercial Mplus, SPSS y Stata. Jamovi, JASP, SPSS y Stata se manejan a través de menús y pueden complementarse con sintaxis, mientras que en Mplus y R se requiere siempre sintaxis. Nuestros siguientes comentarios se refieren a los análisis que pueden manejarse a través de menús o sintaxis, ignorando explícitamente la posibilidad de programar nuevas funciones. La Figura 2 resume esta información.

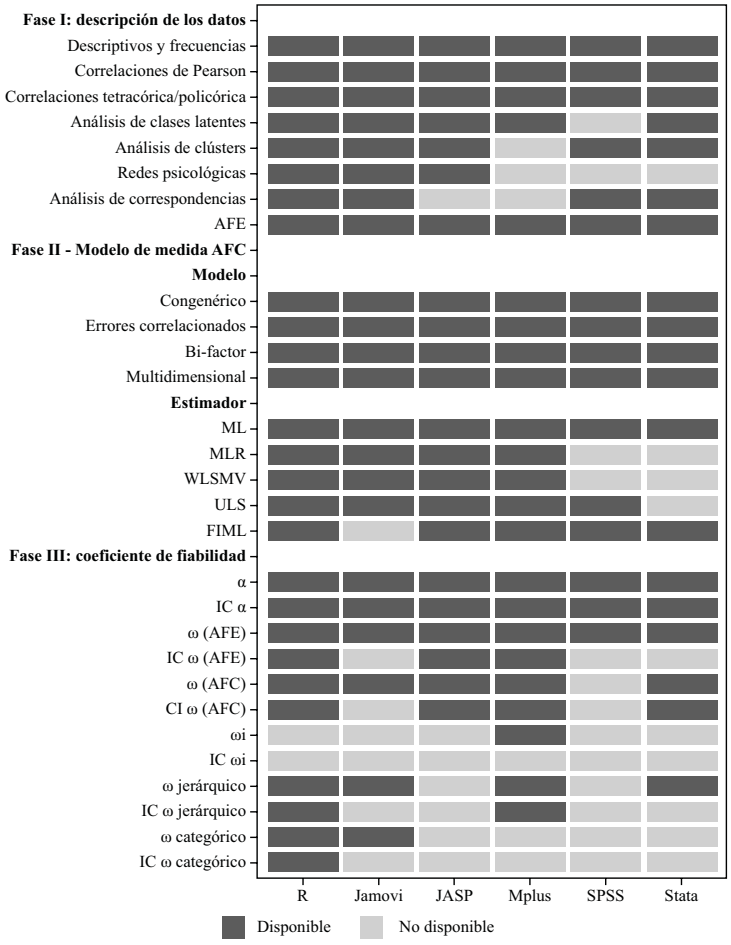

Figura 2.
Comparación de las Posibilidades Analíticas de Varios Programas Informáticos Para Completar el Análisis de Tres Fases Para la Estimación de la Fiabilidad.
Nota: AFE = Análisis factorial exploratorio; AFC = Análisis factorial confirmatorio; ML = Estimador de máxima verosimilitud; MLR = Estimador de máxima verosimilitud robusto; WLSMV = estimador robusto de mínimos cuadrados ponderados con un estadístico $\chi 2$ ajustado para la media y la varianza; ULS = Estimador de mínimos cuadrados no ponderados; FIML = Estimador de máxima verosimilitud con información completa; α = alfa de Cronbach; ω = omega; ω_i = omega con corrección de la especificidad; IC = Intervalo de confianza.

R (R Core Team, 2021). Se puede realizar todos los análisis que hemos sugerido para cada una de las tres fases (es decir, el análisis descriptivo de los ítems, el ajuste del modelo de medida del test y la estimación de la fiabilidad de la consistencia interna de las puntuaciones del test excepto el coeficiente ω_i). La forma más conveniente de obtener resultados en R es adaptar una sintaxis ya hecha. El trabajo de Viladrich et al. (2017) presenta una guía y la sintaxis necesaria para llevar a cabo la Fase 2 y la Fase 3 para tests unidimensionales. Las estimaciones puntuales y de intervalo del coeficiente ω se derivan del AFC. También se proporciona la sintaxis de las estimaciones puntuales y por intervalo del coeficiente α. Complementariamente, Viladrich y Angulo-Brunet (2019) presentan la sintaxis de la Fase 2 y de la Fase 3 para obtener $\omega_{jerárquico}$ basado en un modelo bifactorial confirmatorio. En todas estas sintaxis conviene sustituir la función reliability, obsolescente, por la más actualizada *compRelSEM*. Como ya hemos dicho, si se analiza un modelo multidimensional, la fiabilidad de cada factor puede obtenerse por separado y, por tanto, el procedimiento propuesto en Viladrich et al. (2017) para los test unidimensionales puede aplicarse a cada factor. Además de los análisis confirmatorios, el paquete *psych* (Revelle, 2022; Revelle y Condon, 2019) permite obtener α, ω y $\omega_{jerárquico}$ basados en el modelo bifactorial exploratorio que por defecto asume tres factores menores. Esta opción exploratoria ha sido desaconsejada (e.g., Savalei y Reise, 2019) porque puede proporcionar sobreestimación de la fiabilidad basada en los resultados de un modelo no plausible. Por lo que sabemos, hasta ahora no se ha publicado ninguna sintaxis de R para ω_i.

Jamovi (The Jamovi Project, 2021). A través de menús pueden realizarse todos los análisis propuestos para la Fase 1. Todos los modelos de medida que hemos tratado en la Fase 2 pueden analizarse descargando el módulo complementario semlj (Gallucci y Jentschke, 2021) que instala el menú SEM. En la Fase 3, con el menú SEM, se puede obtener α, y las mismas opciones de α que actualmente ofrece el paquete R psych. Algunas particularidades de este módulo son que no implementa el estimador FIML y que con datos categóricos calcula la versión ordinal de α (Zumbo et al., 2007). Desaconsejamos el uso rutinario del menú *Factor* preinstalado. Aunque ofrece CFA y EFA para datos cuantitativos, y la opción de análisis de fiabilidad calcula los coeficientes α y ω, debe tenerse en cuenta que el valor de ω en la salida solo es correcto para el modelo unidimensional de medidas congenéricas, que es el predeterminado y no puede ser verificado ni modificado por el usuario.

JASP (JASP Team, 2022). Todas las técnicas estadísticas mencionadas en la Fase 1, excepto el análisis de correspondencias, están disponibles en los menús. En la Fase 2, todos los modelos de medida pueden comprobarse con el menú *Factor* si se elige una estrategia exploratoria o con el menú *SEM* si se prefiere una estrategia confirmatoria. Para la Fase 3, el menú *SEM* proporciona la estimación puntual y por intervalo de α y ω. Sin embargo, ω solo es correcto para el modelo unidimensional con estimación de máxima verosimilitud, que es el predeterminado y no puede ser modificado por el usuario.

Mplus (Muthén y Muthén, 2017). Este software comercial ofrece la gama de opciones más amplia para ajustar el modelo de medida (Fase 2) y también permite la primera fase descriptiva limitándose a las técnicas multivariantes que utilizan variables latentes, lo que excluye el análisis de conglomerados, el análisis de redes psicológicas o el análisis de correspondencias múltiples. Una vez más, la opción más conveniente es adaptar sintaxis ya hechas. Viladrich et al. (2019) proporcionan una guía y la sintaxis para ajustar el modelo de medida y estimar la fiabilidad de los test unidimensionales mediante el AFC. Para los modelos bifactoriales confirmatorios, ver Viladrich y Angulo-Brunet (2019), y para los modelos bifactoriales exploratorios ver García-Garzón et al. (2020). Para el cálculo de ω_i puede verse la sintaxis publicada por Sideridis et al. (2019). Por el momento, no hemos encontrado una sintaxis publicada para calcular directamente $\omega_{categórico}$ en Mplus. Existen posibilidades indirectas que incluyen copiar-pegar los valores de salida de Mplus a SAS (Yang y Xia, 2019) o exportar la salida de Mplus a R utilizando la función *mplus2lavaan* del paquete *MplusAutomation* (Hallquist et al., 2022; sintaxis disponible en Viladrich et al., 2019).

IBM SPSS (IBM Corp., 2021). Todas las técnicas estadísticas de la Fase 1 están disponibles a través de menús, excepto las redes psicológicas y el análisis de clases latentes. El comando ampliado *SPSSINC_HETCOR* descargable desde IBM *developerWorks* permite el cálculo de correlaciones tetracóricas y policóricas mediante un paquete R. Otras opciones incluyen la sintaxis de Lorenzo-Seva y Ferrando (2012; 2015). Los modelos de medida de la Fase 2 pueden ajustarse con IBM SPSS Amos (Arbuckle, 2014) un software adicional para la modelización de ecuaciones estructurales con un número limitado de los métodos de estimación que hemos tratado aquí. Para la Fase 3, el comando *reliability* disponible en el módulo básico proporciona la estimación puntual de α, y desde la versión 27.0 también de ω para modelos unidimensionales, que es el modelo por defecto y no puede ser modificado por el usuario. La correlación intraclase denominada consistencia para las medidas promedio, una opción del comando *reliability*, permite obtener la estimación puntual y por intervalo de α.

Stata (StataCorp, 2021). Todas las técnicas estadísticas de la fase descriptiva, excepto las redes psicológicas, están disponibles. El análisis de los modelos de medida se realiza de forma general con la estimación FIML. Para la tercera fase, Viladrich et al. (2019) proporcionan una sintaxis que facilita la estimación puntual y por intervalo de los coeficientes α y ω para los modelos unidimensionales, mientras que Viladrich y Angulo-Brunet (2019) proporcionan una sintaxis para los modelos bifactoriales y $\omega_{jerárquico}$. Hasta donde sabemos, las sintaxis para el cálculo de los coeficientes $\omega_{categórico}$ y ω_i no están disponibles en la actualidad.

Discusión

En este artículo hemos revisado los conocimientos, las prácticas y las soluciones actuales relativas a la estimación de la fiabilidad de las puntuaciones de los test basada en un diseño de consistencia interna. Los principales resultados de nuestra revisión se presentan en forma de un diagrama de flujo destinado a ayudar a los analistas de datos y a los revisores de artículos. El diagrama de flujo facilita la elección razonada del coeficiente de fiabilidad para las puntuaciones obtenidas por suma o promedio de ítems con escalas de respuesta de categorías ordenadas, dicotómicas o politómicas.

Nuestra primera conclusión es que el clásico coeficiente α derivado de la matriz de varianzas-covarianzas entre ítems funciona bien en la mayoría de los casos. Somos más optimistas que Bentler (2021) cuando concluye sobre los usos de α con un lacónico "Eso está bien. Pero eso es todo. Y no es mucho" (p. 866). En nuestra opinión es bastante, al menos en comparación con los usos de ω, su competidor mejor posicionado, aunque no es suficiente porque ninguno de los dos coeficientes proporciona una estimación correcta de la fiabilidad en todos los casos. De hecho, no existe un único coeficiente que cubra esta función para todos los casos (Cho, 2022; Xiao y Hau, 2022).

Consideramos que es bastante debido a la convincente evidencia que apoya un rendimiento similar para α y ω cuando los datos son aproximadamente unidimensionales, las medidas son congenéricas sin cargas factoriales extremas, y las muestras son grandes. El uso de cualquiera de los dos coeficientes sería correcto en este escenario. Y ambos serían incorrectos para modelos de medida con errores correlacionados o ítems que tengan un fuerte componente específico.

Los estudios que comparan α y ω en diferentes condiciones muestran que, en muchos casos, la diferencia entre ambos valores es mínima. En nuestra revisión encontramos que en los estudios de simulación las conclusiones a favor de ω son exageradas ya que, bajo fiabilidades poblacionales razonables desde un punto de vista práctico, la diferencia entre los dos coeficientes se refleja a partir del tercer decimal. Esto se suma a las conclusiones de Savalei y Reise (2019) de que McNeish (2018) exageró la diferencia existente entre los dos coeficientes, y de que las consecuencias de la divergencia a efectos prácticos serían triviales.

Además, el uso de ω conlleva algunos peligros. El más grave se deriva de las decisiones subjetivas que implica el ajuste del modelo de medida. Esta subjetividad puede conducir a resultados mucho más inadecuados que el uso de α así como dificultar su réplica (Davenport et al., 2016; Edwards et al., 2021; Foster, 2021) por no hablar de la mala práctica de seleccionar un modelo de medida ateórico pero estadísticamente ajustado, obtenido a partir de refinamientos basados en los resultados. En nuestra opinión, la mejor manera de enfrentar este peligro es hacer transparentes todas las etapas del análisis, incluida la disponibilidad de la base de datos y la sintaxis utilizada para el análisis.

Así, frente a las propuestas que solo aconsejan la publicación de alguna forma de ω (e.g., Flora, 2020), creemos que α es apropiado en una gran variedad de situaciones. Consideramos que el coeficiente α es sencillo de calcular, comunicar y replicar, y que en los estudios de simulación no difiere de ω hasta el tercer decimal, por lo que tiene una utilidad práctica sin una pérdida sustantiva en el rigor de la estimación de la fiabilidad. Vamos a esperar si en el futuro esta conclusión recibe el soporte de réplicas de estudios de simulación como propone Cho (2022).

Por ahora, la posición más conservadora sería informar de α y ω, como proponen Revelle y Condon (2019). La publicación de α facilitará la comparación directa con otros estudios (de hecho, α sigue siendo el coeficiente de fiabilidad más reportado). Además, la publicación de ω proporcionará una estimación basada en el modelo de medida. Si la diferencia entre los coeficientes α y ω fuera relevante, valdría la pena discutir las razones de esta diferencia.

También hay que decir claramente que ambos coeficientes comparten varias limitaciones. Para empezar, ninguno de ellos es útil para estimar la fiabilidad de la consistencia interna de las puntuaciones derivadas de modelos de medida no lineales o con distribuciones de los residuales que se aparten en gran medida de la normalidad. Para estos casos, en este texto se han tratado los coeficientes derivados de modelos de medida linealizados, como $\omega_{ordinal}$ (Zumbo et al., 2007) y $\omega_{categórico}$ (Green y Yang, 2009), pero los investigadores también deberían considerar los coeficientes derivados de los modelos TRI (Culpepper, 2013; Kim y Feldt, 2010) o la estimación bayesiana aplicable a una amplia variedad de distribuciones exponenciales (Foster, 2020, 2021) que no se han tratado en este texto.

Además, el uso de α y ω se limita a la estimación de la fiabilidad de las puntuaciones obtenidas por suma o promedio de ítems. La generalización de estos coeficientes para estimar la fiabilidad de las puntuaciones factoriales puede verse en Rodriguez et al. (2016b) y en Ferrando y Lorenzo-Seva (2016, 2018). Estos trabajos también abordan otra cuestión todavía más importante, a saber, la discusión sobre la utilidad psicométrica de los coeficientes de fiabilidad en comparación con otros indicadores de la calidad de las puntuaciones factoriales, como la determinación factorial y la varianza común explicada por el factor general. Se trata de una cuestión práctica muy relevante porque en los análisis más habituales con modelos de ecuaciones estructurales, la medida de los constructos latentes no se obtiene por la suma o el promedio de los ítems, sino por la combinación factorial de las respuestas de los ítems. En resumen, aunque las posiciones extremas de las letras α y ω en el alfabeto griego sugieren que se trata de coeficientes situados en las antípodas, las evidencias demuestran que resuelven cuestiones psicométricas bastante parecidas.

Otro punto importante a efectos prácticos es que no hay atajos para calcular α y ω. De hecho, una idea que ha sobrevivido a la discusión de los coeficientes en las últimas décadas es que, sea cual sea el coeficiente que se utilice, la estimación de la fiabilidad de la consistencia interna viene después de probar el modelo de medida. Esta idea está ahora bien establecida y se incluye en textos normativos como el manual de publicación el American Psychological Association (2020) o las directrices de calidad metodológica para el meta-análisis (Prinsen et al., 2018; Sánchez-Meca et al., 2021). Es decir, antes de calcular la fiabilidad de consistencia interna con α u ω, se debe comprobar que un AF de los ítems muestre resultados compatibles con la TCT. Y añadimos que antes se debe comprobar cuál es el tipo adecuado de AF a través de la exploración de los datos. Nuestra visión del análisis como un viaje de tres etapas se alinea con las personas expertas que afirman que no hay formas rápidas de calcular la fiabilidad de la consistencia interna (e.g., Liddell y Kruschke, 2018; Savalei y Reise, 2019) y se aleja del punto de vista de otros expertos que abogan por la difusión de un software específico que conduzca a una aproximación de ω en pocos pasos evitando la evaluación del modelo de medida, como por ejemplo puede hacerse con la macro de Hayes y Coutts (2020) de SPSS. Como han demostrado los estudios de simulación, en la mayoría de los casos, una buena aproximación para ω es simplemente α.

En este ámbito, nuestra contribución específica consiste en señalar que no solo el camino es largo sino que, en sus curvas, los investigadores encontrarán especies tan inesperadas en la psicometría de "apuntar y clicar" como el análisis de conglomerados, la toma de decisiones sobre la relación esperada entre los ítems y los factores, sobre qué partes de la variabilidad de las respuestas se van a considerar varianza verdadera o variancia de error, o qué forma de la distribución de la varianza residual resulta razonable. La recompensa será un profundo conocimiento de sus datos, del grupo humano que ha participado y de la teoría en la que se basa su test.

Otro riesgo es pensar que las demostraciones empíricas sobre la utilidad y eficiencia de ω para datos cuantitativos unidimensionales son generalizables a cualquier otra versión del coeficiente ω como, por ejemplo, está implícito en Flora (2020), en Lai (2021) o en Bentler (2017). En estos trabajos se introducen nuevos coeficientes basados en ω y se suele proporcionar una solución informática para calcularlos. Por un lado, la advertencia de Revelle y Condon (2019) contra la tentación de aplicar fórmulas de fiabilidad a las matrices de correlaciones tetracóricas o policóricas y, por otro, el debate sobre el coeficiente $\omega_{ordinal}$ (Chalmers, 2018; Yang y Green, 2015; Zumbo y Kroc, 2019) nos ha

hecho ser más cautos a la hora de sacar conclusiones en relación con los nuevos coeficientes. Por ello, hemos cambiado de opinión respecto a nuestro trabajo anterior (Viladrich et al., 2017). Actualmente creemos que las generalizaciones matemáticas de ω a nuevas condiciones analíticas deberían acompañarse de investigaciones empíricas comparativas, como la de Yang y Green (2015) o la más reciente de Béland y Falk (2022), que muestren sus ventajas.

En cuanto al software, la implementación del coeficiente α está muy extendida, mientras que la del coeficiente ω es más restringida. Si las características de los datos y del modelo se alinean con la trayectoria sombreada de la Figura 1, la elección del software no será un problema importante para ω y menos aún para α, ya que ambos coeficientes estarán, por lo general, disponibles. A medida que las características de los datos o del modelo se alejan del ideal sombreado en la Figura 1 (e.g., cuando las relaciones no son lineales, algunos ítems presentan errores correlacionados, los datos son ordinales) la necesidad de calcular un tipo particular de ω también requerirá el acceso y el conocimiento de paquetes de software especializados. Queremos advertir contra el uso irreflexivo de programas informáticos bajo el epígrafe *reliabililty* o similar. Algunos de ellos, como la opción *reliability* del menú Factor de Jamovi, el menú *reliabililty* de JASP o la función omega del paquete *psych* de R, proporcionan unos resultados de fiabilidad sin que el usuario tenga control sobre el modelo de medida, mientras que este modelo de medida es de suma importancia, ya que el numerador de ω se basa en las cargas factoriales. Cabe señalar que, por el momento, estas soluciones se basan en AFE unidimensionales que solo serían apropiados cuando los datos presentan las condiciones sombreadas a la izquierda de la Figura 1. Además, por lo general resulta difícil de conseguir la documentación relativa a los métodos subyacentes a una determinada opción del menú, siendo la excepción el bien documentado paquete *psych* (Revelle y Condon, 2019). En su lugar, favorecemos el uso de funciones como *compRelSEM* del paquete *semTools* en R que derivan el cálculo de los coeficientes de fiabilidad a partir de los parámetros estimados al ajustar el modelo de medida. En otras palabras, cuando los datos se apartan de la trayectoria sombreada de la Figura 1, las personas que investigan y revisan solo deberían confiar en las funciones en las que ω es un subproducto de un análisis factorial definido por aquellas que analizaron los datos y no obtenido por defecto en algún paquete estadístico.

En consonancia con lo anterior, cabe hacer algunos comentarios sobre los métodos de meta-análisis de generalización de la fiabilidad. Como hemos dicho repetidas veces, compartimos la indicación de que se debe tener en cuenta el modelo de medida. Sin embargo, una vez ajustado el modelo unidimensional, la agregación de los resultados de fiabilidad se hace sin distinguir entre sus estimadores, ya sean α u ω (Sánchez-Meca et al., 2021) o solo α (Prinsen et al., 2018). Quizás no distinguir entre α y ω_{total} podría ser una buena idea, ya que ambos coeficientes comparten la definición de varianza verdadera y, por tanto, pretenden estimar el mismo parámetro poblacional. Por otro lado, consideramos que los resultados obtenidos con $\omega_{jerárquico}$ o con ω_i no son agregables ni entre sí ni con α o con ω_{total}, ya que la varianza verdadera se define de forma no comparable. Por lo tanto, deben tratarse por separado, como ya es práctica común con otros coeficientes que no comparten con α la definición de varianza verdadera, como el coeficiente de correlación intraclase de acuerdo absoluto (Prinsen et al., 2018). Pensamos que en todos los estudios se debería informar explícitamente de qué parte de la varianza de las respuestas se ha considerado como varianza verdadera. En esta línea, Cho (2022), llega a una conclusión similar, y Scherer y Teo (2020) proponen la solución, más drástica, de realizar meta-análisis de generalización de la fiabilidad sobre la base de las matrices de varianzas-covarianzas y no sobre la base de los coeficientes informados en los estudios primarios. Este tipo de análisis, denominado meta-análisis basado en modelos de ecuaciones estructurales o MASEM, se está desarrollando rápidamente para el estudio de la generalización de la fiabilidad (Sánchez-Meca, 2022).

Pasando al diseño del estudio y al análisis de datos, las personas que investigan deberán superar el marco mental de obtener los datos con una única administración del test para considerar a posteriori cuál es la mejor fórmula para estimar la fiabilidad de la consistencia interna. De hecho, es necesario tener desde el principio claras todas las fuentes de variación para incluirlas en el diseño de la recogida de datos. Por ejemplo, si se quiere medir un constructo conceptualmente amplio con pocos ítems, éstos tendrán una especificidad considerable. Este conocimiento permitirá diseñar la recogida de datos de forma que sea posible estimarla e incluirla como varianza verdadera (Bentler, 2017; Raykov, 2007). O tal vez se pueda optar por incluir algunas fuentes de error en el análisis del modelo de medida como se hace, por ejemplo, en el trabajo de Ferrando y Navarro-González (2021) quienes, utilizando un diseño transversal, proponen un modelo de análisis de datos en el que se estima el error atribuible a cada persona para cuantificar el papel que desempeña en la fiabilidad de un test.

Por novedosas que parezcan estas propuestas, a nuestro juicio, se suman a lo que fue y sigue siendo el objetivo de la TG desde los años cincuenta del siglo pasado. Como hemos dicho, desde esta teoría, el estudio de la fiabilidad se concibe como la identificación y control de las posibles fuentes de error en las puntuaciones de los tests. Los diseños y análisis propuestos por Bentler (2017), Ferrando y Navarro-González (2021), Green y Hershberger (2000) o Raykov (2007) simplemente promueven el control estadístico de las fuentes de error frente al control experimental inicialmente adoptado en la TG.

Pensamos que abordar la cuestión de cómo controlar o, al menos, predecir las posibles fuentes de error ayuda a resolver un hándicap bien conocido de todos los coeficientes de fiabilidad. Estos coeficientes dependen no solo del test, sino también del grupo de personas al que se aplica y del procedimiento de corrección. Según Ellis (2021) una forma de afrontar esta situación es reconocer explícitamente que un mismo test puede tener múltiples fiabilidades. Es decir, al aceptar que no existe una fórmula única para estimar la fiabilidad, y que la mejor fórmula dependerá de lo que se considere error para cada uso previsto de un test, hemos dado el primer paso para admitir que tampoco existe un número fijo de diseños para cubrir este propósito. Para cada uso propuesto de un test será necesario justificar qué evidencias de fiabilidad serían convincentes, tal como recomiendan Muñiz y Fonseca-Pedrero (2019), Ziegler (2020) y, queda reflejado en los distintos grupos de evidencias de fiabilidad recogidos en los *Estándares Para Pruebas Educativas y Psicológicas* (AERA et al., 2014). Este punto de vista conlleva una ampliación manifiesta de los tres diseños clásicos de consistencia interna, fiabilidad test-retest y medidas paralelas, y en consecuencia también de la interpretación de los coeficientes resultantes.

Una última recomendación para editores y revisores sería que, además de valorar la elección del coeficiente, habría que dársela también a los puntos de corte aceptables y a la publicación de los intervalos de confianza. Este mensaje no es nuevo, pero lo repetimos porque parece difícil de aplicar. Aunque han sido muchos los autores que han proporcionado puntos de corte para los coeficientes de fiabilidad (ver, por ejemplo, DeVellis, 2003; Nunnally y Bernstein, 1994, o más recientemente Kalkbrenner, 2021; Taber, 2018), generalmente

basados en opiniones personales (Streiner, 2003), la propuesta de Nunnally y Bernstein (1994) es la más reconocida, y en ella se basan explícitamente, por ejemplo, las recomendaciones del modelo COTAN (Evers et al., 2013). Nunnally y Bernstein (1994) basaron su propuesta en el uso de las puntuaciones de los test y establecieron dos tipos de uso: utilizar las puntuaciones para obtener correlaciones con otras variables o utilizarlas para hacer valoraciones de las personas. En el primer caso establecieron un punto de corte de fiabilidad en .80 para garantizar que la pérdida de fiabilidad en las medidas no provocara una gran atenuación en las correlaciones. Con el fin de obtener medidas muy precisas para el segundo uso, elevaron el valor mínimo de fiabilidad a .90. Sin embargo, estos autores son citados a menudo para justificar valores de fiabilidad de .70, cuando restringen este valor a "las primeras etapas de la investigación predictiva o de validación de constructos" (p. 264). Aunque de sus recomendaciones no se desprende que ninguno de estos valores deba tomarse como punto de referencia absoluto, ni están respaldados por estudios empíricos, muchas personas investigadoras, revisoras y editoras recurren a ellos, especialmente al criterio inferior de .70, como puntos de corte absolutos (Cortina et al., 2020; ver también Lance et al., 2006). También llama la atención que sea habitual proporcionar estimaciones puntuales de los coeficientes, ya sea α u ω, sin el intervalo de confianza del coeficiente como indicador del nivel de precisión de la estimación, lo que debería ser una práctica habitual para las estimaciones muestrales, como se afirma en los textos normativos (Evers et al., 2015; Prinsen et al., 2018; Sánchez-Meca et al., 2021). Hay que tener en cuenta que el valor que debería superar el punto de corte es el límite inferior del intervalo. Una alternativa al tratamiento de la incertidumbre es la que se propone en Pfadt et al. (2022) basada en la estimación bayesiana de estos coeficientes.

En resumen, si se planea estudiar la fiabilidad de consistencia interna de un test, sería recomendable a) organizar la recogida de datos para incluir variables que tengan en cuenta todas las fuentes de error conocidas; b) analizar los datos explorando la completitud, la forma y las relaciones de los datos y evaluando el ajuste del modelo de medida del test; c) informar de la estimación por intervalo de la fiabilidad, utilizando α u otros coeficientes; y d) calibrar su valor en función del uso previsto del test.

Con este trabajo no pretendemos cerrar el debate sobre el uso de los coeficientes de consistencia interna y ni mucho menos sobre la estimación de la fiabilidad. En la actualidad, el debate es tan rico y amplio que abordar todos sus extremos requeriría mucho más espacio que el disponible en este trabajo. Además, como puede observarse por las referencias, se trata de un campo en continuo desarrollo al que habrá que prestar mucha atención en el futuro.

Referencias

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *The Standards for Educational and Psychological Testing*. American Educational Research Association.

Agresti, A. (1996). *An introduction to categorical data analysis*. Wiley.

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). Author. https://doi.org/10.1037/000016S-000

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist, 73*(1), 3–25. https://doi.org/10.1037/amp0000191

Arbuckle, J. L. (2014). Amos (Version 23.0) [Computer software]. IBM SPSS.

Bandalos, D. L. (2021). Item meaning and order as causes of correlated residuals in confirmatory factor analysis. *Structural Equation Modeling, 28*(6), 903–913. https://doi.org/10.1080/10705511.2021.1916395

Béland, S., & Falk, C. F. (2022). A comparison of modern and popular approaches to calculating reliability for dichotomously scored items. *Applied Psychological Measurement, 46*(4), 321–337. https://doi.org/10.1177/01466216221084210

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

Bentler, P. M. (2017). Specificity-enhanced reliability coefficients. *Psychological Methods, 22*(3), 527–540. https://doi.org/10.1037/met0000092

Bentler, P. M. (2021). Alpha, FACTT, and beyond. *Psychometrika, 86*(4), 861–868. https://doi.org/10.1007/s11336-021-09797-8

Bovaird, J. A., & Koziol, N. A. (2012). Measurement models for ordered-categorical indicators. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 495–511). Guilford.

Brennan, R. L. (2001). *Generalizability theory*. Springer.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford.

Chalmers, R. P. (2018). On misconceptions and the limited usefulness of ordinal alpha. *Educational and Psychological Measurement, 78*(6), 1056–1071. https://doi.org/10.1177/0013164417727036

Cho, E. (2022). The accuracy of reliability coefficients: A reanalysis of existing simulations. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000475

Cho, S. J., Shen, J., & Naveiras, M. (2019). Multilevel reliability measures of latent scores within an item response theory framework. *Multivariate Behavioral Research, 54*(6), 856–881. https://doi.org/10.1080/00273171.2019.1596780

Christensen, W. F., Wall, M. M., & Moustaki, I. (2022). Assessing dimensionality in dichotomous items when many subjects have all-zero responses: An example from psychiatry and a solution using mixture models. *Applied Psychological Measurement, 46*(3), 167–184. https://doi.org/10.1177/01466216211066602

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Cortina, J. M., Sheng, Z., Keener, S. K., Keeler, K. R., Grubb, L. K., Schmitt, N., Tonidandel, S., Summerville, K. M., Heggestad, E. D., & Banks, G. C. (2020). From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the Journal of Applied Psychology. *Journal of Applied Psychology, 105*(12), 1351–1381. https://doi.org/10.1037/apl0000815

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

Cronbach, L. J., Nageswari, R., & Gleser, G. C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology, 16*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Culpepper, S. A. (2013). The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. *Applied Psychological Measurement, 37*(3), 201–225. https://doi.org/10.1177/0146621612470210

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. https://doi.org/10.1111/emip.12106

Deng, L., & Chan, W. (2017). Testing the difference between reliability coefficients alpha and omega. *Educational and Psychological Measurement, 77*(2), 185–203. https://doi.org/10.1177/0013164416658325

DeVellis, R. F. (2003). *Scale development. Theory and applications*. Sage.

DiStefano, C., Shi, D., & Morgan, G. B. (2020). Collapsing categories is often more advantageous than modeling sparse data: Investigations in the CFA framework. *Structural Equation Modeling, 28*(2), 237–249. https://doi.org/10.1080/10705511.2020.1803073

Edwards, A. A., Joyner, K. J., & Schatschneider, C. (2021). A simulation study on the performance of different reliability estimation methods. *Educational and Psychological Measurement, 81*(6), 1–29. https://doi.org/10.1177/0013164421994184

Ellis, J. L. (2021). A test can have multiple reliabilities. *Psychometrika, 86*(4), 869–876. https://doi.org/10.1007/s11336-021-09800-2

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*. Guilford.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2015). *COTAN Review System for Evaluating Test Quality*. https://psynip.nl/wp-content/uploads/2022/05/COTAN-review-system-for-evaluating-test-quality.pdf

Evers, A., Muñiz, J., Hagemeister, C., Hstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema, 25*(3), 283–291. https://doi.org/10.7334/psicothema2013.97

Ferrando, P. J., & Lorenzo-seva, U. (2016). A note on improving EAP trait estimation in oblique factor-analytic and item response theory models. *Psicológica, 37*, 235–247.

Ferrando, P. J., & Lorenzo-Seva, U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement, 78(*5), 762–780. https://doi.org/10.1177/0013164417719308

Ferrando, P. J., Lorenzo-seva, U., Hernández-Dorado, A., & Muñiz, J. (2022). Decalogue for the factor analysis of test items. *Psicothema, 34*(1), 7–17. https://doi.org/10.7334/psicothema2021.456

Ferrando, P. J., & Navarro-González, D. (2021). Reliability and external validity of personality test scores: The role of person and item error. *Psicothema, 33*(2), 259–267. https://doi.org/10.7334/psicothema2020.346

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science, 3*(4), 484–501. https://doi.org/10.1177/2515245920951747

Foldnes, N., & Grønneberg, S. (2020). Pernicious polychorics: The impact and detection of underlying non-normality. *Structural Equation Modeling, 27*(4), 525–543. https://doi.org/10.1080/10705511.2019.1673168

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. https://doi.org/10.1080/10705510903203573

Foster, R. C. (2020). A generalized framework for classical test theory. *Journal of Mathematical Psychology, 96*, Article 102330. https://doi.org/10.1016/j.jmp.2020.102330

Foster, R. C. (2021). KR20 and KR21 for some nondichotomous data (It's not just Cronbach's alpha). *Educational and Psychological Measurement, 81*(6), 1172–1202. https://doi.org/10.1177/0013164421992535

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.

Gallucci, M., & Jentschke, S. (2021). *Semlj: Jamovi SEM Analysis* [Computer software]. https://semlj.github.io

García-Garzón, E., Nieto, M. D., Garrido, L. E., & Abad, F. J. (2020). Bi-factor exploratory structural equation modeling done right: using the slidapp application. *Psicothema, 32*(4), 607–614. https://doi.org/10.7334/psicothema2020.179

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLoS ONE, 12*(6), 1–26. https://doi.org/10.1371/journal.pone.0174035

Goodboy, A. K., & Martin, M. M. (2020). Omega over alpha for reliability estimation of unidimensional communication measures. *Annals of the International Communication Association, 44*(4), 422–439. https://doi.org/10.1080/23808985.2020.1846135

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. https://doi.org/10.1177/0013164406288165

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling, 7*(2), 251–270. https://doi.org/10.1207/S15328007SEM0702_6

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. https://doi.org/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. https://doi.org/10.1111/emip.12100

Greenacre, M. (2017). *Correspondence analysis in practice* (3rd ed.). Chapman & Hall. https://doi.org/10.1201/9781315369983

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. https://doi.org/10.1027/1614-2241/a000052

Gulliksen, H. (1950). *Theory of mental tests*. Wiley.

Hallquist, M., Willey, J., van Lissa, C., & Morillo, D. (2022). *MplusAutomation: an R package for facilitating large-scale latent variable analyses in Mplus* (1.1.0) [Computer software]. https://michaelhallquist.github.io/MplusAutomation/

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But… *Communication Methods and Measures, 14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G., & Elosua, P. (2016). Revisión del modelo para evaluar la calidad de los tests utilizados en España [Assessing the quality of tests in spain: revision of the spanish test review model]. *Papeles Del Psicólogo, 37*(3), 192–197.

Hoekstra, R., Vugteveen, J., Warrens, M. J., & Kruyen, P. M. (2019). An empirical analysis of alleged misunderstandings of coefficient alpha. *International Journal of Social Research Methodology, 22*(4), 351–364. https://doi.org/10.1080/13645579.2018.1547523

Hox, J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications*. Routledge.

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science, 3*(2), 166–184. https://doi.org/10.1177/2515245919882903

IBM Corp. (2021). *IBM SPSS Statistics for Windows, Version 28.0* (28.0) [Computer software]. IBM Corp.

JASP Team. (2022). *JASP (Jeffreys's Amazing Statistics Program)* (0.16.2) [Computer software]. https://jasp-stats.org/

Kalkbrenner, M. T. (2021). Alpha, omega, and H internal consistency reliability estimates: Reviewing these options and when to use them. Counseling Outcome Research and Evaluation, Published. *Advance online publication*, 1–12. https://doi.org/10.1080/21501378.2021.1940118

Kim, S., Lu, Z., & Cohen, A. S. (2020). Reliability for tests with items having different numbers of ordered categories. *Applied Psychological Measurement, 44*(2), 137–149. https://doi.org/10.1177/0146621619835498

Kim, Seonghoon, & 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. https://doi.org/10.1007/s12564-009-9062-8

Komperda, R., Pentecost, T. C., & Barbera, J. (2018). Moving beyond alpha: A primer on alternative sources of single-administration reliability evidence for quantitative chemistry education research. *Journal of Chemical Education, 95*(9), 1477–1491. https://doi.org/10.1021/acs.jchemed.8b00220

Lai, M. H. C. (2021). Composite reliability of multilevel data: It's about observed scores and construct meanings. *Psychological Methods, 26*(1), 90–102. https://doi.org/10.1037/met0000287

Lance, C., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods, 9*(2), 202–220. https://doi.org/10.1177/1094428105284919

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

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 [The exploratory factor analysis of the items: a practical guide, revised and updated]. *Anales de Psicología, 30*(3), 1151–1169. https://doi.org/10.6018/analesps.30.3.199361

Lorenzo-Seva, U., & Ferrando, P. J. (2012). TETRA-COM: A comprehensive SPSS program for estimating the tetrachoric correlation. *Behavior Research Methods, 44*(4), 1191–1196. https://doi.org/10.3758/s13428-012-0200-6

Lorenzo-Seva, U., & Ferrando, P. J. (2015). POLYMAT-C: A comprehensive SPSS program for computing the polychoric correlation matrix. *Behavior Research Methods, 47*(3), 884–889. https://doi.org/10.3758/s13428-014-0511-x

McCrae, R. R. (2015). A more nuanced view of reliability: specificity in the trait hierarchy. *Personality and Social Psychology Review, 19*(2), 97–112. https://doi.org/10.1177/1088868314541857

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

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://dx.doi.org/10.1037/met0000144

Muñiz, J. (2018). *Introducción a las Teorías Psicométricas* [Introduction to Psychometric Theories]. Pirámide.

Muñiz, J., & Fonseca-Pedrero, E. (2019). Ten steps for test development. *Psicothema, 31*(1), 7–16. https://doi.org/10.7334/psicothema2018.291

Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2016). *Regression and mediation analysis using Mplus*. Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide (8th edition)*. Muthén & Muthén.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory. McGraw-Hill.

Olvera, O. L., Kroc, E., & Zumbo, B. D. (2020). The role of item distributions on reliability estimation: the case of Cronbach's coefficient alpha. *Educational and Psychological Measurement, 80*(5), 825–846. https://doi.org/10.1177/0013164420903770

Oosterwijk, P. R., van der Ark, L. A., & Sijtsma, K. (2019). Using confidence intervals for assessing reliability of real tests. *Assessment, 26*(7), 1207–1216. https://doi.org/10.1177/1073191117737375

Pfadt, J. M., van den Bergh, D., Klaas, S., Moshagen, M., & Wagenmakers, E.-J. (2022). Bayesian estimation of single-test reliability coefficients bayesian estimation of single-test reliability coefficients. 57(4), 620–641. *Multivariate Behavioural Research, 57*(4). https://doi.org/10.1080/00273171.2021.1891855

Pons, J., Viladrich, C., & Ramis, Y. (2017). Examining the big three of coping in adolescent athletes using network analysis. *Revista de Psicologia Del Deporte, 26*, 68–74.

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1147–1157. https://doi.org/10.1007/s11136-018-1798-3

R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

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

Raykov, T. (1997b). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32*(4), 329–353. https://doi.org/10.1207/s15327906mbr3204_2

Raykov, T. (2001). Bias of coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25*(1), 69–76. https://doi.org/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. https://doi.org/10.1207/s15328007sem1103

Raykov, T. (2007). Reliability of multiple-component measuring instruments: Improved evaluation in repeated measure designs. *British Journal of Mathematical and Statistical Psychology, 60*(1), 119–136. https://doi.org/10.1348/000711006X100464

Raykov, T., Anthony, J. C., & Menold, N. (2022). On the importance of coefficient alpha for measurement research: loading equality is not necessary for alpha's utility as a scale reliability index. *Educational and Psychological Measurement*. Advance online publication. https://doi.org/10.1177/00131644221104972

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. https://doi.org/10.1080/10705511003659417

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. https://doi.org/10.1177/0013164414526039

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

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement, 79*(1), 200–210. https://doi.org/10.1177/0013164417725127

Raykov, T., Marcoulides, G. A., Harrison, M., & Menold, N. (2019). Multiple-component measurement instruments in heterogeneous populations: Is there

a single coefficient alpha? *Educational and Psychological Measurement, 79*(2), 399–412. https://doi.org/10.1177/0013164417733305

Revelle, W. (2022). *psych: Procedures for personality and pscychological research* (2.2.5) [Computer software]. https://personality-project.org/r/psych/

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: *A tutorial. Psychological Assessment, 31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

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. https://doi.org/10.1037/a0029315

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249

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

Sánchez-Meca, J. (2022, july 20). *Meta-análisis de generalización de la fiabilidad* [Reliability generalization Meta-analysis][Simposium]. XVII Congreso de Metodología de Las Ciencias Sociales y de La Salud.

Sánchez-Meca, Julio, Marín-Martínez, F., López-López, J. A., Núñez-Núñez, R. M., Rubio-Aparicio, M., López-García, J. J., López-Pina, J. A., Blázquez-Rincón, D. M., López-Ibáñez, C., & López-Nicolás, R. (2021). Improving the reporting quality of reliability generalization meta-analyses: The REGEMA checklist. *Research Synthesis Methods, 12*(4), 516–536. https://doi.org/10.1002/jrsm.1487

Savalei, V., & Reise, S. P. (2019). Don't forget the model in your model-based reliability coefficients: A reply to McNeish (2018). *Collabra: Psychology, 5*(1), 36. https://doi.org/10.1525/collabra.247

Savalei, V., & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology, 66*(2), 201–223. https://doi.org/10.1111/j.2044-8317.2012.02049.x

Scherer, R., & Teo, T. (2020). A tutorial on the meta-analytic structural equation modeling of reliability coefficients. *Psychologycal Methods, 25*(6), 747–775. https://doi.org/10.1037/14262-002

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350–353.

Shi, D., Lee, T., Fairchild, A. J., & Maydeu-Olivares, A. (2020). Fitting ordinal factor analysis models with missing data: A comparison between pairwise deletion and multiple imputation. *Educational and Psychological Measurement, 80*(1), 41–66. https://doi.org/10.1177/0013164419845039

Sideridis, G. D., Tsaousis, I., & Al-Sadaawi, A. (2019). An application of reliability estimation in longitudinal designs through modeling item-specific error variance. *Educational and Psychological Measurement, 79*(6), 1038–1063. https://doi.org/10.1177/0013164419843162

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

Sijtsma, K., & Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: discussing lower bounds and correlated errors. *Psychometrika, 86*, 843–860. https://doi.org/10.1007/s11336-021-09789-8

Slaney, K. L., Tkatchouk, M., Gabriel, S. M., & Maraun, M. D. (2009). Psychometric assessment and reporting practices: Incongruence between theory and practice. *Journal of Psychoeducational Assessment, 27*(6), 465–476. https://doi.org/10.1177/0734282909335781

StataCorp. (2021). *Stata statistical software.* (Release 17) [Computer software]. StataCorp LLC. https://www.stata.com/

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99–103.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

The jamovi project (2021). *The jamovi project (v1.6)* [Computer software]. https://www.jamovi.org

Thurstone, L. L. (1947). *Multiple factor analysis.* University of Chicago Press.

Viladrich, C., y Angulo-Brunet, A. (2019). *Reliability of Essentially Unidimensional Measures Derived From Bifactor Modeling With R, Mplus and Stata.* [Data set and syntax]. Universitat Autònoma de Barcelona. https://ddd.uab.cat/record/205936

Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Annals of Psychology, 33*(3), 755–782. https://doi.org/10.6018/analesps.33.3.268401

Viladrich, C., Angulo-Brunet, A., & Doval, E. (2019). *Mplus and stata tools to calculate the internal consistency reliability coefficients alpha and omega* [Data set and syntax]. Universitat Autònoma de Barcelona. https://ddd.uab.cat/record/205870

Weijters, B., Geuens, M., & Schillewaert, N. (2009). The proximity effect: The role of inter-item distance on reverse-item bias. *International Journal of Research in Marketing, 26*(1), 2–12. https://doi.org/10.1016/j.ijresmar.2008.09.003

Xiao, L., & Hau, K.T. (2022). Performance of coefficient alpha and its alternatives: Effects of different types of non-normality. *Educational and Psychological Measurement*. Advance online publication. https://doi.org/10.1177/00131644221088240

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st Century? *Journal of Psychoeducational Assessment, 29*(4), 377–392. https://doi.org/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. https://doi.org/10.1027/1614-2241/a000087

Yang, Y., & Xia, Y. (2019). Categorical omega with small sample sizes via bayesian estimation: An alternative to frequentist estimators. *Educational and Psychological Measurement, 79*(1), 19–39. https://doi.org/10.1177/0013164417752008

Ziegler, M. (2020). Psychological test adaptation and development – How papers are structured and why. *Psychological Test Adaptation and Development*. Advance online publication. https://doi.org/10.1027/2698-1866/a000002

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. https://doi.org/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. https://doi.org/10.1107/S0907444909031205

Zumbo, B. D., & Kroc, E. (2019). A measurement is a choice and Stevens' scales of measurement do not help make it: A response to Chalmers. *Educational and Psychological Measurement, 79*(6), 1184–1197. https://doi.org/10.1177/0013164419844305