Psychologia. Avances de la disciplina
ISSN: 1900-2386
psychologia@usb.edu.co
Universidad de San Buenaventura
Colombia

Revisión de los criterios para validez convergente estimada a través de la Varianza Media Extraída

Moral de la Rubia, José

Revisión de los criterios para validez convergente estimada a través de la Varianza Media Extraída
Psychologia. Avances de la disciplina, vol. 13, núm. 2, 2019
Universidad de San Buenaventura, Colombia
Disponible en: https://www.redalyc.org/articulo.oa?id=297261276003

Revisión de los criterios para validez convergente estimada a través de la Varianza Media Extraída

Review of the criteria for convergent validity estimated through the Extracted Average Variance

José Moral de la Rubia jose_moral@hotmail.com
Universidad Autónoma de Nuevo León, México

Psychologia. Avances de la disciplina, vol. 13, núm. 2, 2019

Universidad de San Buenaventura, Colombia

Recepción: 16 Enero 2019
Aprobación: 08 Agosto 2019

Redalyc: https://www.redalyc.org/articulo.oa?id=297261276003

Resumen: En el contexto del modelamiento de ecuaciones estructurales se manejan los conceptos de validez convergente y confiabilidad compuesta aplicados a los modelos de medida con ítems congenéricos. Este estudio metodológico tiene como objetivo revisar los puntos de corte estipulados para la Varianza Media Extraída (VME) cuando se usa para establecer la validez convergente de un modelo de medida. Por una parte, se consideró la gran semejanza entre los conceptos de validez convergente y confiabilidad por consistencia interna, al usarse los pesos de medida estandarizados para su estimación. Por otra parte, se analizó la relación entre el número de ítems del factor, la VME y los coeficientes omega y H. A efectos de simplificación, se manejó un peso de medida homogéneo en las demostraciones. Se observó un efecto muy grande del número de ítems, provocando un deterioro en la VME, aun manteniendo constante el mismo nivel de confiabilidad (coeficientes omega y H), y más acusado a mayor nivel de confiabilidad. Se concluye que valores de VME < .50 pueden reflejar niveles aceptables de validez convergente, dependiendo del número de ítems, si incluyen como criterios complementarios: pesos de medida estandarizados ≥ .50 y coeficientes omega y H ≥ .70.

Palabras clave: validez convergente, confiabilidad compuesta, confiabilidad de constructo, análisis factorial confirmatorio, modelo de medida.

Abstract: In the context of the structural equation modeling, the concepts of convergent validity and composite reliability applied to the measurement models with congeneric items are used. This methodological study aims to review the stipulated cut-off points for the Extracted Average Variance (AVE) when it is used to establish the convergent validity of a measurement model. On the one hand, the great similarity between the concepts of convergent validity and internal consistency reliability, when standardized measurement weights are used for their estimation, was considered. On the other hand, the relationships among the number of factor items, the AVE and the omega and H coefficients were analyzed. For simplification purposes, a homogeneous measurement weight was used in the demonstrations. A very large effect of the number of items was observed, causing a deterioration in the AVE, while still maintaining the same level of reliability (omega and H coefficients), and more pronounced at a higher level of reliability. It is concluded that AVE values < .50 may reflect acceptable levels of convergent validity depending on the number of items, if the following criteria are included as complementary criteria: standardized measurement weights ≥ .50, and coefficients omega and H ≥ .70.

Keywords: convergent validity, composite reliability, construct reliability, confirmatory factor analysis, measurement model.

Introducción

Conceptos psicométricos básicos

Se inicia el manuscrito con la definición de los conceptos de psicometría, instrumento de medida, medición y escala de medida para encuadrar los conceptos de confiabilidad y validez.

La psicometría es el campo de estudio concerniente a las teorías y técnicas de medición en psicología. En psicometría, el término de instrumento de medida hace referencia a un test o cuestionario formado por unas instrucciones verbales generales y unos ítems, reactivos o preguntas. Los ítems pueden ser verbales, lógicos, numéricos o materiales manipulables y requerir respuestas orales, escritas o ejecuciones de conductas observables (Kline, 2015). A su vez, el formato de respuesta en los ítems puede ser cerrado, y usualmente homogéneo, o puede ser abierto. En este último caso, habría una codificación establecida para categorizar las respuestas, como en las escalas de conducta y las pruebas proyectivas (Piotrowski, 2015). Unos instrumentos de medida requieren reportar conductas, ya sea públicamente observables o internas (sensaciones, deseos, motivos, emociones o pensamientos), las cuales son características de la persona; como es el caso de los test de personalidad, motivación, estado emocional o actitud. Otros instrumentos requieren resolver problemas, retener información, comprender información o mostrar conocimiento; como es el caso de los test de inteligencia, capacidad o funcionalidad (McClimans, Brown & Canoc, 2017).

Las opciones de respuesta de los ítems del test se transforman en números bajo una regla estipulada. A este proceso de asignar números a las respuestas de los participantes se le denomina medición (Loewenthal & Lewis, 2018).

La escala de medida de un ítem hace referencia a la correspondencia entre las propiedades empíricas de las distintas modalidades de respuesta y las propiedades algebraicas y operaciones aritméticas admisibles de los números asignados. Se distinguen cuatro tipos de escala de medida: nominal, ordinal, de intervalo o de razón (Wu & Leung, 2017).

La escala de medida nominal asigna símbolos numéricos a los distintos atributos o modalidades cualitativas de respuesta. Estos números permiten contar frecuencias y contingencias, pero no admiten operaciones aritméticas ni poseen propiedades algebraicas. Usualmente es de tipo dicotómico, como por ejemplo, 0 = "no" y 1 = "sí" ante la ejecución de una conducta. Otro ejemplo sería: 0 = "conducta impropia del rasgo medido" y 1 = "conducta propia del rasgo".

Una escala ordinal asigna números ordinales a las distintas categorías ordenables de las modalidades de respuesta. Estos números poseen propiedades de ordenación y conteo, así como posibilidad de aritmética ordinal de suma, multiplicación y exponenciación. Es la más usada en escalas de actitud y emociones e inventarios de personalidad y motivaciones. Un ejemplo sería: 1 = "nunca", 2 = "a veces", 3 = "con

frecuencia", 4 = "con mucha frecuencia" y 5 = "siempre" ante la frecuencia de una conducta.

Las escalas de medida también pueden ser de intervalo (por ejemplo, distancia desde el origen de la línea hasta el punto marcado en un continuo que va a 0 "nada" a 10 "insoportable" al evaluar el malestar provocado por un síntoma) o de razón (por ejemplo "tiempo de respuesta medido en milisegundos" o "número de veces que se repite una conducta en un intervalo de tiempo"). Estas dos escalas asignan números naturales, enteros, racionales, irracionales o reales a las distintas cantidades o magnitudes en que se presenta la respuesta. Estos números admiten operaciones aritméticas y poseen propiedades algebraicas, poseyendo los números reales las propiedades y posibilidades aritméticas más amplias o inclusivas. En la escala de intervalo, no se puede fijar un cero absoluto (ausencia del rasgo) y la unidad de medida solo permite medir distancias relativas entre puntos (intervalos), y en la escala de razón, se puede fijar el cero absoluto y las mediciones son absolutas desde este punto de origen.

La puntuación en el test se obtiene por la suma simple o el promedio de las puntuaciones en los ítems (Kline, 2015; Streiner, Norman & Cairney, 2015). Las puntuaciones en el test pueden ser interpretadas en términos absolutos o relativos. En el primer caso, se usan el contenido de las categorías de respuesta. En el segundo caso, se usan como normas interpretativas las estimaciones de los parámetros de la distribución teórica a la que se aproximan las puntuaciones en una población, o se usan las estimaciones de los cuantiles poblacionales en caso de que la función de distribución empírica no siga ninguna función teórica reconocible (Loewenthal & Lewis, 2018). Usualmente, las puntuaciones en los test de actitud, emociones, motivaciones y personalidad están en una escala de medida de intervalo (Wu & Leung, 2017).

Confiabilidad y validez del instrumento de medida

Desde la teoría clásica de test, la confiabilidad de un instrumento de medida hace referencia a su capacidad de obtener mediciones con error mínimo (Jabrayilov, Emons & Sijtsma, 2016). Se puede establecer a través de tres estrategias: confiabilidad temporal o correlaciones al aplicar el mismo instrumento en dos o más momentos distintos; confiabilidad interjuez o correlaciones entre las mediciones obtenidas por distintos evaluadores al aplicar el instrumento a los mismos participantes, y confiabilidad por consistencia interna o correlación entre los distintos ítems o formas paralelas del test (Kline, 2015).

Desde la teoría clásica de test, la validez hace referencia a si un instrumento realmente mide lo que afirma evaluar (Moses, 2017). Tradicionalmente se distinguen tres tipos de validez: de contenido o en qué medida los ítems cubren el dominio de contenido del constructo; de criterio o correlación con otra medida confiable y válida que mide el mismo constructo u otro muy afín, y de constructo o grado en que se confirman las relaciones esperadas bajo la teoría y definiciones que sustentan al constructo. La validez de constructo incluye tanto el

contraste del modelo estructural como la aportación acumulativa de evidencias de validez concurrentes, predictivas y retrodictivas (Furr, 2017).

Se considera que, una vez demostrada la validez de contenido, criterial y de constructo, queda establecida la validez de un instrumento de medida (Moses, 2017). Este enfoque tradicional de tres tipos de validez empezó a ser revisado en la década de 1970 (Furr, 2017). Por una parte, Cronbach (1971) enfatizó que las evidencias de validez ofrecen apoyo a las inferencias hechas con un instrumento en un contexto dado, con una muestra dada. Por otra parte, Messick (1975) concibió la validez como un conjunto de evidencias que ayudan a establecer la relación entre las puntuaciones en un instrumento y el constructo medido, esto es, el grado en que la evidencia y la teoría apoyan las interpretaciones de las puntuaciones en el test desde los usos propuestos para el mismo.

Ambos enfoques desplazaron la atención hacia la aplicación del instrumento en contextos concretos frente al planteamiento inicial más general o universalista, en el cual la confiabilidad y validez una vez probadas ya quedan establecidas. De este nuevo enfoque surge la necesidad de validar y baremar el instrumento de medida en los distintos contextos en que se aplica, como distintos contextos culturales (Loewenthal & Lewis, 2018).

Con el desarrollo del análisis factorial confirmatorio (Jöreskog, 1978; Jöreskog, Olsson, & Wallentin, 2016), no solo vino una nueva estrategia para aportar evidencias de validez de constructo (Bollen, 1989;), sino también nuevas propuestas para el cálculo de la confiabilidad a través de la consistencia interna sin asumir que exista una covarianza homogénea entre las puntuaciones verdaderas y los errores de medida de los ítems (tau-equivalencia). Por ejemplo, se tienen las propuestas de Jöreskog (1971) y Werts, Rock, Linn y Jöreskog (1978) sobre el coeficiente de confiabilidad compuesta.

La validez convergente del factor

Al usarse análisis factorial confirmatorio cada factor posee unos indicadores, y se considera como requisito necesario demostrar que los indicadores propuestos miden dicho factor (Jöreskog, 1978). De aquí surge el concepto de la validez convergente de un modelo de medida, cuyo antecedente es la validez a través de matrices de correlación multirasgo-multimétodo introducida por Campbell y Fiske a finales de la década de 1950 (Jöreskog et al., 2016).

¿Qué es la validez convergente de un modelo de medida o factor? Dado un factor con n indicadores o ítems, este tipo de validez hace referencia al grado de certeza que se tiene en que los indicadores propuestos miden una misma variable latente o factor. Al preguntar si un modelo de medida posee validez convergente se pretende averiguar si el constructo es adecuadamente medido por los indicadores propuestos (Cheung & Wang, 2017).

Fornell y Larcker (1981) propusieron la VME calculada desde los pesos de medida estandarizados para evaluar la validez convergente del modelo de medida. Dado un factor o variable latente con k indicadores, la (ecuación 1) es igual al promedio de los k pesos de medida estandarizados elevados al cuadrado ( ), los cuales son estimados a partir de una muestra de n participantes, usando uno de los métodos desarrollados para modelamiento de ecuaciones estructurales, como máxima verosimilitud o mínimos cuadrados no ponderados (Fornell & Larcker, 1981).

$$\overline{VME} = \frac{\sum_{i=1}^{k} \hat{\lambda}_i^{\,2}}{\sum_{i=1}^{k} \hat{\lambda}_i^{\,2} + \sum_{i=1}^{k} \hat{\sigma}_{\hat{\varepsilon}_i}^2} = \frac{\sum_{i=1}^{k} \hat{\lambda}_i^{\,2}}{\sum_{i=1}^{k} \hat{\lambda}_i^{\,2} + \sum_{i=1}^{k} \left(1 - \hat{\lambda}_i^{\,2}\right)} = \frac{\sum_{i=1}^{k} \hat{\lambda}_i^{\,2}}{k} = \overline{\lambda^2}$$

Ecuación 1. Varianza Media Extraída

Ecuación 1.

Varianza Media Extraída

Fornell y Larcker (1981) fijaron el criterio de que un factor, con independencia de su número de indicadores, debe explicar más del 50 % de la varianza de los mismos para que se pueda considerar que posee validez convergente (nivel aceptable), e indicaron que, de preferencia, debería explicar más del 70 % (nivel bueno). El argumento es que la varianza atribuible al factor sea mayor que la no atribuible.

Hair, Black, Babin y Anderson (2010), al requisito de una VME mayor que .5, le añadieron pesos de medidas estandarizadas de al menos .5 en todos los indicadores y una confiabilidad compuesta de al menos .7. Tras un estudio de simulación, Cheung y Wang (2017) también recomendaron una VME no significativamente menor que .50 y unos pesos de medidas estandarizados no significativamente menores que .50 en todos los indicadores.

La confiabilidad compuesta o de constructo

Actualmente, como señalan Green y Yang (2015), se recomienda el uso del coeficiente de confiabilidad compuesta ( ) para evaluar la confiabilidad por consistencia interna (ecuación 2). Este índice también es conocido como coeficiente ρ de Jöreskog (Jöreskog, 1971, 1978) o coeficiente ω de McDonald (1999). Se ideó para modelos de medida congenéricos y así poder superar la limitación del coeficiente alfa de Cronbach que requiere que los ítems sean esencialmente tau-equivalentes (Cho & Kim, 2015).

La varianza verdadera se estima a través del cuadrado de la suma de los pesos de medida estandarizados. Para obtener la varianza del test, a la varianza verdadera se le suma la varianza del error de medida. La varianza del error de medida se calcula por la suma de la varianza de los residuos o suma de cuadrados de los complementos de los pesos de medida estandarizados. El cociente entre la varianza verdadera y la varianza del test proporciona el coeficiente de confiabilidad (Green & Yang, 2015).

$$\widehat{CR} = \hat{\rho} = \hat{\omega} = \frac{\left(\sum_{i=1}^{k} \hat{\lambda}_i\right)^2}{\left(\sum_{i=1}^{k} \hat{\lambda}_i\right)^2 + \sum_{i=1}^{k} \hat{\sigma}_{e_i}^2} = \frac{\left(\sum_{i=1}^{k} \hat{\lambda}_i\right)^2}{\left(\sum_{i=1}^{k} \hat{\lambda}_i\right)^2 + \sum_{i=1}^{k} \left(1 - \hat{\lambda}_i^2\right)}$$

Ecuación 2. Confiabilidad compuesta, coeficiente rho o coeficiente omega

Ecuación 2.

Confiabilidad compuesta, coeficiente rho o coeficiente omega

Se ha estipulado que valores entre .70 y .79 en el coeficiente omega reflejan niveles de confiabilidad por consistencia interna aceptables al indicar que al menos el 70 % de la varianza de las mediciones o puntuaciones empíricas en el test están sin error. A su vez, unos valores entre .80 y .89 se consideran buenos, y mayores o iguales que .90 se juzgan excelentes (Cho & Kim, 2015; Green & Yang, 2015; Viladrich, Angulo-Brunet & Doval, 2017). No obstante, se señala que valores mayores que .94 pueden indicar la presencia de ítems redundantes (Kline, 2015).

Una variante del coeficiente de confiabilidad compuesta es el coeficiente de confiabilidad congenérica ( ) de Cho (2016). En su fórmula (Ecuación 3), el denominador del coeficiente de confiabilidad es la varianza del test, lo que lo hace que sea un índice más comparable con otros coeficientes de confiabilidad basados en la teoría clásica de test y desarrollados bajo el supuesto de ítems esencialmente tau-equivalentes (Cho, 2016).

$$\hat{\rho}_C = \frac{\left(\sum_{i=1}^{k} \hat{\lambda}_i\right)^2}{\hat{s}_x^2}$$

[~~Ecuación 3. Del coeficiente~~ de confiabilidad congenérica]

Al plantear que los ítems son esencialmente tau-equivalentes se parten de unos supuestos recogidos dentro de la teoría clásica de test o del error de medida (Pakzad & Alaeddini, 2017). Estos supuestos son: 1) la puntuación observada en un ítem es la suma de dos puntuaciones independientes, una puntuación verdadera y un error de medida; 2) la media de los errores de medida del ítem es cero; 3) los errores de medida de dos ítems de la escala son independientes, y 4) las puntuaciones verdaderas de un ítem y el error de medida del otro ítem también son independientes (Jabrayilov et al., 2016).

Bajo estos supuestos, se dice que tres o más ítems son paralelos si sus puntuaciones verdaderas y las varianzas de sus errores de medida son equivalentes; esto implica que las medias, varianzas, covarianzas y correlaciones de las puntuaciones observadas son estadísticamente homogéneas. Tres o más ítems son tau-equivalentes si sus puntuaciones verdaderas son equivalentes, aunque las varianzas de sus errores de medida no lo sean; esto implica que las medias y covarianzas de

las puntuaciones observadas son estadísticamente equivalentes, pero las varianzas y correlaciones pueden ser heterogéneas. Tres o más ítems son esencialmente tau-equivalentes si sus puntuaciones verdaderas son combinación lineal unas de otras bajo operaciones de adición o sustracción, aunque las varianzas de sus errores sean heterogéneas; esto implica que las covarianzas entre las puntuaciones observadas son estadísticamente homogéneas, aunque las medias, varianzas y correlaciones de las puntuaciones observadas pueden ser heterogéneas (Jabrayilov et al., 2016; Pakzad & Alaeddini, 2017).

Entre las nuevas técnicas para evaluar la confiabilidad por consistencia interna, también se tiene el coeficiente H de Hancock y Müeller (2001) o confiabilidad de la combinación lineal óptima entre ítems (Ecuación 4). Este coeficiente se propone como una estimación del límite superior de la confiabilidad por consistencia interna de un test sin requerir el supuesto de tau-equivalencia (Domínguez-Lara, 2016), como el coeficiente ω, que se basa en los pesos de medida estandarizados, además, no requiere que los ítems estén puntuados en el mismo sentido, como sí ocurre con el coeficiente α y el coeficiente ω al estar los pesos de medida elevados al cuadrado (McNeish, 2018). Se requiere un valor de al menos .70 para reflejar confiabilidad (Domínguez-Lara, 2016; Hancock & Müeller, 2001).

$$\hat{H} = \frac{\sum_{i=1}^{k} \frac{\hat{\lambda}_i^2}{1-\hat{\lambda}_i^2}}{1+\sum_{i=1}^{k} \frac{\hat{\lambda}_i^2}{1-\hat{\lambda}_i^2}} = \frac{1}{1+\frac{1}{\sum_{i=1}^{k} \frac{\hat{\lambda}_i^2}{1-\hat{\lambda}_i^2}}}$$

[Ecuación 4. Coeficiente de confiabilidad H]

Diferencia y semejanza entre validez convergente y confiabilidad de constructo

La diferencia entre ambos conceptos radica en que la validez convergente intenta estimar en qué proporción la varianza de los indicadores de un factor es explicada por una fuente atribuible (el factor) y no por fuentes no atribuibles (los errores). La confiabilidad compuesta o de constructo pretende separar la varianza del factor medida sin error (varianza verdadera) y con error (varianza del error de medida). La varianza del test es la suma de la varianza verdadera y la del error de medida. En un principio son dos conceptos próximos o afines, pero claramente distinguibles.

La semejanza fuerte aparece cuando las estimaciones se hacen desde los pesos de medidas estandarizados que el factor tiene sobre los ítems bajo el supuesto de que los errores de medidas son independientes (Figura 1). La varianza verdadera se obtiene al sumar los pesos de medida y elevar esta suma al cuadrado. La varianza del error de medida se obtiene al

sumar los complementos de la varianza explicada por el factor en cada ítem. Esta varianza explicada es el peso de medida elevado al cuadrado. El promedio de estos pesos al cuadrado proporciona la varianza atribuible al factor y su complemento, la varianza no atribuible. En este punto, la semejanza los hace muy parecidos y complementarios, como ya lo reconocen investigadores como Green y Yang (2015), Hair et al. (2010) y Rodríguez, Reise y Haviland (2016).

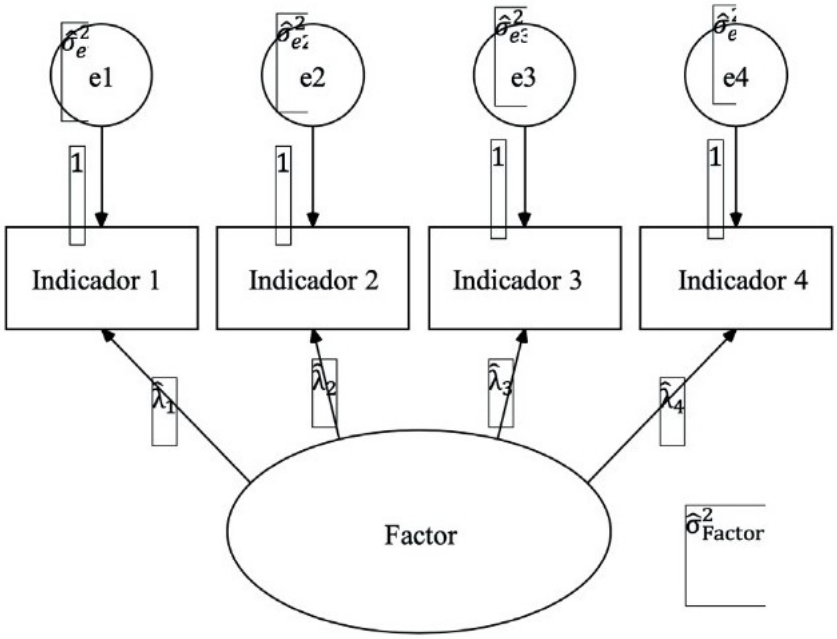

Figura 1. Modelo de un factor con dos grados de libertad o diferencia entre los 10 momentos muestrales (seis covarianzas y cuatro varianzas) y los ocho parámetros a estimar (las cuatro varianzas de residuos de medida y los cuatro pesos de medida tras fijar a un valor unitario la varianza del factor y los pesos de los residuos).

Figura 1

Modelo de un factor con dos grados de libertad o diferencia entre los 10 momentos muestrales (seis covarianzas y cuatro varianzas) y los ocho parámetros a estimar (las cuatro varianzas de residuos de medida y los cuatro pesos de medida tras fijar a un valor unitario la varianza del factor y los pesos de los residuos).

Se podría decir que ambos conceptos pretenden estimar una varianza atribuible a un modelo de medida especificado con k indicadores. Dicho en otras palabras, ambos conceptos plantean responder a la pregunta de si el modelo de medida de un constructo posee una variabilidad compartida o varianza verdadera (sin error) suficiente para ser una buena medición. Desde esta similitud se podría reconsiderar el valor crítico para la VME desde los valores críticos propuestos para el coeficiente de confiabilidad compuesta (coeficiente ω) y de constructo (coeficiente H).

A parir de lo previamente argumentado se plantea un estudio metodológico que tiene como objetivo revisar el punto de corte estipulado para la VME cuando se usa para establecer la validez convergente de un modelo de medida.

Método

En los análisis y demostraciones se consideró que todos los ítems presentan la misma correlación con el factor, es decir, sus pesos de medida estandarizados son equivalentes (Pakzad, & Alaeddini, 2017). Esta homogeneidad de pesos de medida ( ) se adoptó como un planteamiento simplificado, aunque no sea requerida por los coeficientes ω (Ecuación 2) y H (Ecuación 4), ya que permite derivar una fórmula sencilla para obtener el peso de medida mínimo para un nivel de confiabilidad dado en los coeficientes ω o H (.70 aceptable, .80 bueno o .90 excelente). Además, esta situación de estricta tau-equivalencia entre los ítems simplifica el cálculo de la VME, que aparece en la ecuación 1, ya que la media de una constante es la constante:

$$\widehat{VME} = \hat{\lambda}_i^2$$

[Ecuación 5. Cálculo de Varianza Media Extraída con pesos de medida homogéneos]

Se tomó como indicadores de confiabilidad por consistencia interna a los coeficientes ω (Ecuación 2) y H (Ecuación 4). Ambos coeficientes no son invariantes con respecto al número de indicadores (Abdelmoula, Chakroun & Fathi, 2015). Para comprobar este hecho se calculó la correlación entre el número de indicadores y el valor del coeficiente. Este cálculo se hizo a través del coeficiente de correlación producto-momento de Pearson (r). Se interpretó que un valor absoluto de $|r| < .10$ muestra una fuerza de asociación trivial, entre .10 y .29 pequeña, entre .30 y .49 media, entre .50 y .69 grande, entre .70 y .89 muy grande y ≥ .90 perfecta o unitaria (Schober, Boer & Schwarte, 2018).

Para analizar la relación entre el número de indicadores, VME y coeficientes ω y H, se calculó qué valor debe tener el peso de medida estandarizado homogéneo en el factor ( ) para alcanzarse los valores mínimos estipulados como confiabilidad aceptable (= = .70), buena (.80) o excelente (.90), variando el número de indicadores (k). Se partió del número mínimo de indicadores para un factor dentro de una estructura de dos o más factores (k = 3) y se llegó hasta un número de indicadores considerado muy grande para un factor (k = 30). Obtenido el peso de medida estandarizado homogéneo, la del factor se calculó elevando dicho peso al cuadrado. A continuación, se comparó la fuerza de asociación entre el número de indicadores y la entre las tres condiciones de confiabilidad para evidenciar el deterioro de la validez convergente con el incremento de los indicadores y que este deterioro es más fuerte a un mayor nivel de confiabilidad. Las comparaciones se hicieron por la prueba Z de Steiger (1980) en un contraste a dos colas y un nivel de

significación de .05. Se aplicó la corrección de Bonferroni (1936) para comparaciones múltiples (α = .05/6), con lo que se requirió un valor de probabilidad < .0083 para rechazar la hipótesis nula de equivalencia entre las correlaciones.

Por otra parte, también variando el número de indicadores (k de 3 a 30), se calcularon los valores de los coeficientes ω y H que corresponden a un peso de medida estandarizado homogéneo alto (= .50 con una = .25 baja) y a dos pesos de medida estandarizados homogéneos muy altos (= .71 con una = .50 aceptable y = .84 y una = .70 buena). A continuación, se comparó la fuerza de asociación entre el número de indicadores y los valores en los coeficientes ω o H entre las tres condiciones de validez convergente para evidenciar un comportamiento opuesto al anterior. Las comparaciones se hicieron por la prueba Z de Steiger (1980) y se usó la corrección de Bonferroni (1936) para comparaciones múltiples. Todos los cálculos se ejecutaron con el programa Excel 2013.

Fornell y Larcker (1981), con independencia del número de indicadores y del valor del coeficiente de confiabilidad compuesta, estipularon como puntos de corte para establecer validez convergente: ≤ .50 mala, > .50 y < .70 aceptable y ≥ .70 buena. Desde la relación evidenciada entre el número de indicadores y los valores de , y , se hace la propuesta para modificar los puntos de corte en AVE estipulados para validez convergente. Esta propuesta se basa en considerar el número de indicadores, un valor mínimo para el peso de medida y un valor mínimo para los coeficientes o .

Resultados

Valores del peso de medida estandarizado homogéneo para alcanzar valores mínimos de confiabilidad aceptable, buena o excelente variando el número de indicadores del factor

¿Qué peso medida estandarizado se requeriría para alcanzar valores de .70, .80 y .90 en los coeficientes ω y H? Considerando la correlación entre el ítem y el factor o peso de medida estandarizado constante, se tendría para el coeficiente ω de McDonald (1999):

$$\widehat{\omega} = \frac{\left(\sum_{i=1}^{k} \hat{\lambda}_i\right)^2}{\left(\sum_{i=1}^{k} \hat{\lambda}_i\right)^2 + \sum_{i=1}^{k}\left(1 - \hat{\lambda}_i^{\,2}\right)} = \frac{\left(k * \hat{\lambda}_i\right)^2}{\left(k * \hat{\lambda}_i\right)^2 + k + k * \hat{\lambda}_i^{\,2}} = \frac{k^2 * \hat{\lambda}_i^{\,2}}{k^2 * \hat{\lambda}_i^{\,2} + 1 - k * \hat{\lambda}_i^{\,2}}$$

$$\frac{k^2 * \hat{\lambda}_i^{\,2}}{k * \left(k * \hat{\lambda}_i^{\,2} + 1 - \hat{\lambda}_i^{\,2}\right)} = \frac{k * \hat{\lambda}_i^{\,2}}{k * \hat{\lambda}_i^{\,2} + 1 - \hat{\lambda}_i^{\,2}} = \frac{k * \hat{\lambda}_i^{\,2}}{1 + (k-1) * \hat{\lambda}_i^{\,2}}$$

$$\widehat{\omega} * \left(1 + (k-1) * \hat{\lambda}_i^{\,2}\right) = k * \hat{\lambda}_i^{\,2}$$

[Ecuación 6. Peso de medida homogéneo ante un valor dado de confiabilidad compuesta]

Por ejemplo, para tres indicadores, k = 3 (Tabla 1), aplicando las ecuaciones 6 y 5:

Por ejemplo, para cuatro indicadores, k = 4 (Tabla 1), aplicando las ecuaciones 6 y 5:

Por ejemplo, para cinco indicadores, k = 5 (Tabla 1), aplicando las ecuaciones 6 y 5:

Con el coeficiente H de Hancock y Müeller (2001) resultaría igual:

$$\hat{H} * \left(1 - \hat{\lambda}_i^{\,2} * (k-1)\right) = k * \hat{\lambda}_i^{\,2}$$

$$\hat{H} + \hat{H} * k * \hat{\lambda}_i^{\,2} - \hat{H} * \hat{\lambda}_i^{\,2} = k * \hat{\lambda}_i^{\,2}$$

$$\hat{H} = k * \hat{\lambda}_i^{\,2} - \hat{H} * k * \hat{\lambda}_i^{\,2} + \hat{H} * \hat{\lambda}_i^{\,2}$$

$$\hat{H} = \hat{\lambda}_i^{\,2} * \left(k - \hat{H} * k + \hat{H}\right)$$

$$\hat{\lambda}_i^{\,2} = \frac{\hat{H}}{k - \hat{H} * k + \hat{\omega}} = \frac{\hat{\omega}}{k - \hat{H} * (k-1)}$$

$$\hat{\lambda}_i = \sqrt{\frac{\hat{H}}{k - \hat{H} * (k-1)}}$$

[Ecuación 7. Peso de medida homogéneo ante un valor dado del coeficiente H]

Tabla 1 Valores para los pesos de medidas estandarizados ( ) y la varianza media extraída ( ), variando el número de indicadores del factor para obtener el valor mínimo de confiabilidad por consistencia interna, considerado aceptable (= = .70), bueno (= = .80) y excelente (= = .90) ante una situación de pesos de medida estandarizados homogéneos.

Tabla 1

k	= = .70		= = .80		= = .90	
3	.661	.438	.756	.571	.866	.750
4	.607	.368	.707	.500	.832	.692
5	.564	.318	.667	.444	.802	.643
6	.529	.280	.632	.400	.775	.600
7	.500	.250	.603	.364	.750	.563
8	.475	.226	.577	.333	.728	.529
9	.454	.206	.555	.308	.707	.500
10	.435	.189	.535	.286	.688	.474
11	.418	.175	.516	.267	.671	.450
12	.403	.163	.500	.250	.655	.429
13	.390	.152	.485	.235	.640	.409
14	.378	.143	.471	.222	.626	.391
15	.367	.135	.459	.211	.612	.375
16	.357	.127	.447	.200	.600	.360
17	.347	.121	.436	.190	.588	.346
18	.339	.115	.426	.182	.577	.333
19	.331	.109	.417	.174	.567	.321
20	.323	.104	.408	.167	.557	.310
21	.316	.100	.400	.160	.548	.300
22	.310	.096	.392	.154	.539	.290
23	.303	.092	.385	.148	.530	.281
24	.298	.089	.378	.143	.522	.273
25	.292	.085	.371	.138	.514	.265
26	.287	.082	.365	.133	.507	.257
27	.282	.080	.359	.129	.500	.250
28	.277	.077	.354	.125	.493	.243
29	.273	.074	.348	.121	.487	.237
30	.269	.072	.343	.118	.480	.231

Valores para los pesos de medidas estandarizados ( ) y la varianza media extraída ( ), variando el número de indicadores del factor para obtener el valor mínimo de confiabilidad por consistencia interna, considerado aceptable (= = .70), bueno (= = .80) y excelente (= = .90) ante una situación de pesos de medida estandarizados homogéneos.
Nota. k = número de indicadores del factor. Estadísticos obtenidos en una muestra de n participantes: = coeficiente omega de McDonald (1999), = coeficiente H de Hancock y Müeller (2001), = varianza media extraída y = peso de medida estandarizado (varianza del factor fijada a uno, media del factor fijada a 0 y suma de la carga factorial al cuadrado y la varianza del residuo de medida igualada a 1 en análisis factor confirmatorio) o correlación del ítem con el factor (peso en la matriz estructural en análisis factorial exploratorio). En negrilla, se destaca cuando se cumple: ≥ .50, o H ≥ .70 y ≥ .25.

Relación entre el número de indicadores del factor y la validez convergente

Desde los datos de la Tabla 1 se obtiene una correlación negativa entre el número de indicadores del factor (k) y el valor del indicador de validez convergente ( ). La fuerza de asociación de estas correlaciones varía de muy grande a perfecta. Cuando = = .70, la correlación entre k y es de -.89 (varianza compartida del 79.1 %). Cuando = = .80, la correlación entre k y es de -.91 (varianza compartida del 83.6 %). Cuando = = .90, la correlación entre k y es de -.95 (varianza compartida del 90.5 %).

Consecuentemente, al incrementar el número de indicadores y mantener constante la confiabilidad, el valor de decae bajo una relación lineal casi perfecta. Al comparar estos coeficientes de correlación por la prueba Z de Steiger, esta fuerza de asociación es significativamente más grande (más negativa) a medida que mejora el nivel de confiabilidad: Z = 4.60, p < .001 al comparar los valores de entre la primera y segunda condición de confiabilidad; Z = 4.91, p < .001 al comparar los valores de entre la primera y la tercera condición de confiabilidad; así como Z = 4.84, p < .001 al comparar los valores de entre la segunda y la tercera. Por tanto, el decremento del valor de con el incremento de indicadores empeora al ser mayor el nivel de confiabilidad.

Valores del coeficiente ω y H ante pesos de medida estandarizados homogéneos de .50, .71 y .84 variando el número de indicadores del factor

Al mantener constantes las correlaciones homogéneas entre los ítems y el factor se puede observar que los coeficientes y son más altos cuanto mayor es el número de indicadores (Tabla 2). Por ejemplo, con un peso de medida estandarizado de .50, el cual es considerado alto en el análisis factorial (Bollen, 1989; Hair et al, 2010; Wu & Leung, 2017), pero que proporciona una VME insuficiente, = .25 (Fornell & Larcker, 1981), la confiabilidad con seis indicadores sería cuestionable, = = .67 < .70 (Hancock & Müeller, 2001; McDonald, 1999); pero con siete indicadores sube a aceptable, = = .70, con 12 indicadores a buena, = = .80 y con 26 indicadores a excelente, = = .90 (Tabla 2).

El número mínimo de indicadores para un factor considerado como admisible es tres en el caso de un modelo de dos o más factores (Jöreskog et al., 2016). Ante un modelo de un factor se requieren cuatro indicadores para que el modelo esté sobreidentificado y permita no solo estimar los parámetros, sino también contrastar el ajuste a los datos (Figura 1). Con tres indicadores con pesos de medida de .66, se logra que los coeficientes y tomen un valor de .70, lo que implica una de .44. Con pesos de medida de .71 se logra una de .50 y los coeficientes y serían de .75, lo que corresponde a un valor de consistencia interna también aceptable. Con pesos de .76 y una de .57, ambos coeficientes subirían a un valor de consistencia interna bueno, = = .80 (Tabla 2).

Entre cuatro y cinco se considera un número medio de indicadores y seis o más, alto (Perry, Nicholls, Clough & Crust, 2015). Con cuatro indicadores con pesos de medida de .61 se logra que los coeficientes ω y H tomen un valor de .70, lo que implica una de .37; con pesos de medida de .71 y de .50 y los coeficientes y sería de .80, lo que corresponde a un valor de consistencia interna bueno. Con cinco indicadores con pesos de medida de .56, se logra que los coeficientes ω y H tomen un valor de .70, lo que implica una VME de .32; con pesos de medida de .71 y de .50 y los coeficientes y sería de .83, lo que corresponde a un valor de consistencia interna bueno. Con seis indicadores con pesos de medida de .53 se logra que los coeficientes ω y H tomen un valor de .70, lo que implica una VME de .28; con pesos de medida de .71 y de .50 y los coeficientes y sería de .86,

lo que corresponde a un valor de consistencia interna bueno. Con siete indicadores con pesos de medida de .50 se logra que los coeficientes ω y H tomen un valor de .70, lo que implica una VME de .25; con pesos de medida de .71 y de .50 y los coeficientes y sería de .88, lo que corresponde a un valor de consistencia interna bueno (Tabla 2).

Más de 20 ítems se suelen considerar un número excesivo de indicadores para un factor (Van der Eijk & Rose, 2015). Con 21 indicadores o más con cargas menores que .60 y valores de menores o iguales que .30, se logran valores de consistencia interna excelentes (y ≥ .90) (Tabla 2).

Tabla 2 Valores de los coeficientes omega y H variando el número de indicadores por factor ante la situación de pesos de medida estandarizados homogéneos.

Tabla 2

k	Valor de =		
	= .25 y = .50	= .50 y = .71	= .70 y =.84
3	.500	.750	.875
4	.571	.800	.903
5	.625	.833	.921
6	.667	.857	.933
7	.700	.875	.942
8	.727	.889	.949
9	.750	.900	.955
10	.769	.909	.959
11	.786	.917	.963
12	.800	.923	.966
13	.813	.929	.968
14	.824	.933	.970
15	.833	.938	.972
16	.842	.941	.974
17	.850	.944	.975
18	.857	.947	.977
19	.864	.950	.978
20	.870	.952	.979
21	.875	.955	.980
22	.880	.957	.981
23	.885	.958	.982
24	.889	.960	.982
25	.893	.962	.983
26	.897	.963	.984
27	.900	.964	.984
28	.903	.966	.985
29	.906	.967	.985
30	.909	.968	.986

Valores de los coeficientes omega y H variando el número de indicadores por factor ante la situación de pesos de medida estandarizados homogéneos.
Nota k = número de indicadores del factor. Estadísticos obtenidos en una muestra de n participantes: = coeficiente omega de McDonald (1999), = coeficiente H de Hancock y Müeller (2001), = varianza media extraída y = peso de medida estandarizado (varianza del factor fijada a uno, media del factor fijada a 0 y suma de la carga factorial al cuadrado y la varianza del residuo de medida igualada a 1 en análisis factor confirmatorio) o correlación del ítem con el factor (peso en la matriz estructural en análisis factorial exploratorio).

Relación entre el número de indicadores del factor y la confiabilidad

Desde los datos de la Tabla 2 se obtiene una correlación positiva entre el número de indicadores y el valor del coeficiente de confiabilidad. La fuerza de asociación de estas correlaciones varía de muy grande a perfecta. Cuando = .25, la correlación entre k y los coeficientes o es de .90 (varianza compartida del 81.1 %). Cuando = .50, la correlación entre k y los coeficientes o es de .86 (varianza compartida del 73.8 %). Cuando = .70, la correlación entre k y los coeficientes o es de .84 (varianza compartida del 70.5 %). Al incrementar el número de indicadores y mantener constante el nivel de validez convergente, el coeficiente de confiabilidad se incrementa bajo una relación lineal casi perfecta. Al comparar estos coeficientes de correlación por la prueba Z de Steiger, esta fuerza de asociación fue significativamente menor a medida que mejora el nivel de validez convergente: $Z = 4.37$, $p < .001$ al comparar los valores de los coeficientes o entre la primera y segunda condición de validez convergente, $Z = 4.48$, $p < .001$ al comparar los valores de los coeficientes o entre la primera y la tercera condición de validez convergente, así como $Z = 4.24$, $p < .001$ al comparar los valores de los coeficientes o entre la segunda y la tercera. Por tanto, el incremento de la confiabilidad con el incremento de indicadores es menor al ser mayor el nivel de validez convergente.

Discusión y conclusiones

El número de ítems del factor tiene un efecto muy fuerte sobre la y los coeficientes y , lo cual no puede ser ignorado. Si el número de ítems aumenta, se deteriora la validez convergente medida a través del ; por el contrario, la confiabilidad (medida por el coeficiente o) aumenta. Además, este deterioro en la validez convergente se manifiesta al mantener constante la confiabilidad y es más acusado a un mayor nivel de confiabilidad. Esto implica que el número de indicadores debe contemplarse al valorar la VME como indicador de validez convergente. Cabe señalar que si los valores de son relativos al número de ítems, y el cálculo de este estadístico depende solo de los pesos de medida, entonces sería importante fijar un valor mínimo de peso de medida. Este problema ya ha sido anteriormente planteado y como valor mínimo se ha propuesto .50 (Hair et al., 2010; Wu & Leung, 2017).

Al establecerse la validez convergente, el promedio de varianza que el factor comparte con los ítems es insuficiente porque no estima adecuadamente el error (Jöreskog, 1971). De ahí que es necesario incluir la división entre varianza verdadera y error de medida sin requerir que los ítems sean esencialmente tau-equivalentes o paralelos (Hair et al., 2010). Precisamente, esta separación es proporcionada de una forma más exacta por los coeficientes y . (Hancock & Müeller, 2001; McDonald, 1999), y un mínimo aceptable de proporción de varianza verdadera puede ser .70, como se viene tradicionalmente estipulando (Cho & Kim, 2015).

Tomando como criterio que los pesos de medida sean mayores o iguales que .50, el coeficiente o sea mayor o igual que .70 y la no sea menor que .25, una de .44 podría ser considerada aceptable para tres indicadores, de .37 para cuatro, .32 para cinco, .28 para seis y .25 para siete, que son las situaciones más comunes de número de ítems por variable latente en modelamiento de ecuaciones estructurales. Entre siete y nueve es necesario subir el valor mínimo de y a .75, entre 10 y 12 a .80 para que la cumpla con el requisito de pesos de medida estandarizados mayores o iguales que .50. Estos valores de , que son menores que .50, reflejarían un nivel aceptable de validez convergente o varianza atribuible a un modelo de medida especificado, valores entre .50 y .70 mostrarían un nivel de validez convergente bueno y mayores que .70 un nivel muy bueno. Consecuentemente, se está proponiendo una interpretación del VME más relativa que la inicialmente planteada por Fornell y Larcker (1981).

La situación de ítems paralelos (correlaciones homogéneas entre los ítems) en el presente estudio se tomó a efectos de simplicidad demostrativa y de cálculo. Por tanto, los señalamientos hechos aplican perfectamente a la confiabilidad por consistencia interna calculada por el coeficiente alfa de Cronbach. No obstante, ni el coeficiente omega ni el coeficiente H requieren este supuesto. En consecuencia, se podría retomar como un peso de medida promedio.

Limitaciones

En el presente estudio no se manipuló el tamaño muestral. Se asumió un tamaño adecuado. ¿Cuál es un tamaño muestral adecuado? No existe una única regla sobre el tamaño muestral en análisis factorial confirmatorio, aunque hay autores, como Kline (2015) y Byrne (2016), que sugieren que sería entre 200 y 400 participantes.

Wolf, Harrington, Clark y Miller (2013) realizaron un estudio con una metodología de simulación. A través del procedimiento de Monte Carlo se generaron 10 000 simulaciones aleatorias de modelos factoriales con datos continuos normalmente distribuidos. En estos modelos se varió el tamaño muestral (n de 20 a 500 con incrementos de 10), la carga factorial homogénea entre los indicadores (= .50, .65 y .80), el número de indicadores por factor (k = 3 ó 4, 6 y 8), el número de factores (1, 2 y 3) y la correlación homogénea entre los factores (r = .30 y .50). Las estimaciones se hicieron por el método de máxima verosimilitud.

En el modelo de un factor, estos investigadores hallaron que con un tamaño muestral de 200 participantes con cuatro indicadores con pesos de .50 se alcanzaba una potencia ≥ .99, no existían problemas de convergencia o soluciones inadmisibles y los sesgos o diferencias entre el parámetro estimado y el valor muestral eran menores que .05 en todos los casos. Estas condiciones de potencia ≥ .99, soluciones convergentes y admisibles y sesgos < .05 se lograron con una muestra de 120 con seis indicadores y de 100 con ocho indicadores. Al ser = .65, estas condiciones se alcanzaron con una n = 100 si el número de indicadores por factores k es 4, con una n = 60 si k = 6 y con una n = 50 si k = 8.

En un modelo de dos factores correlacionados con tres indicadores con cargas de .50 y correlación entre ambos factores de .30, se requiere una n = 400 para lograr una potencia ≥ .85, soluciones convergentes y admisibles, así como estimaciones un sesgo < .05 en las 10 000 simulaciones aleatorias. El tamaño muestral bajaría a 200 con seis indicadores y 130 con ocho indicadores para lograr estos mismos resultados de potencia, convergencia/admisibilidad y sesgo.

Consecuentemente, cuanto menor es el número de factores, mayor el número de indicadores por factor, mayor el tamaño del efecto del factor sobre sus indicadores y mayor la correlación entre los factores, el tamaño de muestra requerido disminuye, pudiendo ser menor que 100 (Kyriazos, 2018). Se podría sugerir que con por lo menos seis indicadores con cargas de al menos .70, n ≥ 100; entre 3 y 5 indicadores con cargas de al menos .70, n ≥ 150. Con al menos seis indicadores con cargas entre .50 y 70 el tamaño muestral debería ser al menos de 200, y entre tres y cinco indicadores con cargas entre .50 y 70 debería ser al menos de 400 (Kyriazos, 2018; Wolf et al., 2013).

Sugerencias para futuras investigaciones

En futuros estudios basados en técnicas de simulación de datos se podría retomar la relación entre el número de ítems y la confiabilidad de constructo desde los promedios de pesos de medida heterogéneos. Cabe señalar que la mayoría de los estudios de simulación se han hecho con el método de máxima verosimilitud y variables continuas normalmente distribuidas, cuando esta no es una situación común en psicometría. En el caso de ítems tipo Likert (que tienen una escala de medida ordinal), la matriz de correlación más adecuada sería la de correlación policórica, y los métodos para optimizar la función de discrepancia serían mínimos cuadrados ponderados robustos, mínimos cuadrados diagonalmente ponderados, mínimos cuadrados no ponderados y mínimos cuadrados simples (Byrne, 2016; Jöreskog et al., 2016; Morata & Holgado, 2013). Un estudio de simulación con este tipo de datos y estos métodos sería más adecuado para la práctica psicométrica.

El requisito de un > .50 puede motivar la omisión de la validez convergente de los factores en los procesos de validación de instrumentos de medida, debido a que es un criterio difícil de satisfacer. De ahí que la presente propuesta de criterios más flexibles para un nivel aceptable de validez convergente (por ejemplo, ≥ .37, ≥ .50 y o ≥ .70 con cuatro indicadores) puede facilitar su inclusión. Para la validez discriminante entre dos factores también se ha propuesto una alternativa al criterio de Fornell y Larcker (1981) de una varianza compartida menor que el AVE de cada factor. Esta alternativa es el criterio de una proporción heterorasgo-monorasgo de las correlaciones entre los ítems de los factores menor o igual que .85 (Henseler, Ringle, & Sarstedt, 2015), el cual puede facilitar el cumplimiento de la propiedad de validez discriminante en modelos de factores correlacionados.

Referencias

Abdelmoula, M., Chakroun, W., & Fathi, A. (2015). The effect of sample size and the number of items on reliability coefficients: Alpha and rho: A meta-analysis. *International Journal of Numerical Methods and Applications, 13*(1), 1-20. doi: 10.17654/IJNMAMar2015_001_020

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley & Sons. doi: 10.1002/9781118619179

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Florencia: Reale Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Byrne, B. M. (2016). *Structural equation modelling with AMOS: Basic concepts, applications, and programming* (3. ed.). New York, NY: Routledge.

Cheung, G. W., & Wang, C. (2017). Current approaches for assessing convergent and discriminant validity with SEM: issues and solutions. *Academy of Management Proceedings, 2017*(1), 12706. doi: 10.5465/AMBPP.2017.12706abstract

Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods, 19*(4), 651-682. doi: 10.1177/1094428116656239

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: well-known but poorly understood. *Organizational Research Methods, 18*(2), 207-230. doi: 10.1177/1094428114555994

Domínguez-Lara, S. (2016). Evaluación de la confiabilidad del constructo mediante el Coeficiente H: breve revisión conceptual y aplicaciones. *Psychologia: Avances en la disciplina, 10*(2), 87-94. doi: 10.21500/19002386.2134

Fornell, C., & Larcker, D. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39-50. doi: 10.2307/3151312

Furr, R. M. (2017). *Psychometrics: an introduction* (3. ed.). New York, NY: Sage Publications.

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 Practices, 34*(4), 14-20. doi: 10.1111/emip.12100

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis: A global perspective* (7. ed.). Upper Saddle River, NJ: Prentice Hall.

Hancock, G. R., & Müeller, R. O. (2001). Rethinking construct reliability within latent variable systems. En R. Cudeck, S. du Toit & D. Sörbom (Eds.), *Structural equation modeling: Present and future—A Festschrift in honor of Karl Jöreskog* (pp. 195-216). Lincolnwood, IL: Scientific Software Internal.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science, 43*(1), 115-135. doi: 10.1007/s11747-014-0403-8

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement, 40*(8), 559-572. doi: 10.1177/0146621616664046

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*(2), 109-133. doi: 10.1007/BF02291393

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices: models, theory, psychometrics, and practice. *Psychometrika, 43*(4), 443-477. doi: 10.1007/BF02293808

Jöreskog, K. G., Olsson, U. H., & Wallentin, F. Y. (2016). Confirmatory factor analysis (CFA). In *Multivariate analysis with LISREL. Springer series in statistics* (pp. 283-339). Switzerland: Springer. doi: 10.1007/978-3-319-33153-9_7

Kline, P. (2015). *A handbook of test construction (psychology revivals). Introduction to psychometric design*. London: Routledge. doi: 10.4324/9781315695990

Kyriazos, T. A. (2018). Applied psychometrics: sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology, 9*, 2207-2230. doi: 10.4236/psych.2018.98126

Loewenthal, K., & Lewis, C. A. (2018). *An introduction to psychological tests and scales* (2. ed.). London: Taylor & Francis Group. doi: 10.4324/9781315782980

McClimans, L., Brown, J., & Canoc, S. (2017). Clinical outcome measurement: Models, theory, psychometrics and practice. *Studies in History and Philosophy of Science Part A, 65-66*, 67-73. doi: 10.1016/j.shpsa.2017.06.004

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

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

Messick, S. (1975). The standard program: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955-956. doi: 10.1037/0003-066X.30.10.955

Morata, M. A., & Holgado, F. P. (2013). Construct validity of Likert scales through confirmatory factor analysis: a simulation study comparing different methods of estimation based on Pearson and polychoric correlations. *International Journal of Social Science Studies, 1*(1), 54-61. doi: 10.11114/ijsss.v1i1.27

Moses, T. (2017). Psychometric contributions: focus on test scores. En R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment the methodological, psychological and policy contributions of ETS* (pp. 79-132). New York, NY: Springer. doi: 10.1007/978-3-319-58689-2_3

Pakzad, R., & Alaeddini, F. (2017). Misuse and misconception of Cronbach's alpha coefficient as an index of internal consistency of measuring tools. *Iranian Journal of Epidemiology, 12*(4), 64-71. Recuperado de http://irje.tums.ac.ir/article-1-5622-en.html

Perry, J. L., Nicholls, A. R., Clough, P. J., & Crust, L. (2015). Assessing model fit: caveats and recommendations for confirmatory factor analysis and exploratory structural equation modeling. *Measurement*

in Physical Education and Exercise Science, 19(1), 12-21. doi: 10.1080/1091367X.2014.952370

Piotrowski, C. (2015). Projective techniques usage worldwide: A review of applied settings 1995-2015. *Journal of the Indian Academy of Applied Psychology, 41*(3), 9-19. URL: https://www.academia.edu/26014801/Projective_Techniques_Usage_Worldwide_A_Review_of_Applied_Settings_1995-2015

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

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia, 126*(5), 1763-1768. doi: 10.1213/ANE.0000000000002864

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245-251. doi: 10.1037/0033-2909.87.2.245

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: a practical guide to their development and use* (5. ed.). Oxford: Oxford University Press. doi: 10.1093/med/9780199685219.001.0001

Van der Eijk, C., & Rose, J. (2015). Risky business: factor analysis of survey data - assessing the probability of incorrect dimensionalisation. *PLoS One, 10*(3), e0118900. doi: 10.1371/journal.pone.0118900

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. doi: 10.6018/analesps.33.3.268401

Werts, C. E., Rock, D. R., Linn, R. L., & Jöreskog, K. G. (1978). A General Method of Estimating the Reliability of a Composite. *Educational and Psychological Measurement, 38*(4), 933-938. doi: 10.1177/001316447803800412

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and Psychological Measurement, 76*(6), 913-934. doi: 10.1177/0013164413495237.

Wu, H., & Leung, S. O. (2017). Can Likert scales be treated as interval scales? A simulation study. *Journal of Social Service Research, 43*(4), 527-532. doi: 10.1080/01488376.2017.1329775