Psicología y Salud, Vol. 27, Núm. 1: 5-18, enero-junio de 2017

Recomendaciones sobre los procedimientos de construcción y validación de instrumentos y escalas de medición en la psicología de la salud

Recommendations about procedures for construction and validation of scales in health psychology

Roberto Lagunes Córdoba[1]

RESUMEN

Aunque existe una gran cantidad de escalas e instrumentos para medir variables psicológicas, las considerables diferencias que hay entre las poblaciones y la variedad de temas que se estudian en la psicología de la salud hacen necesaria la creación de nuevos instrumentos. Desafortunadamente, los procedimientos seguidos son, en muchos casos, inadecuados, y algunos de ellos se desaconsejan en la literatura psicométrica reciente. En esta revisión se presenta un panorama actualizado de dichos procedimientos, empezando por una breve discusión de la naturaleza de la medición psicológica y la manera en que la misma condiciona la medición de las variables y la construcción de instrumentos para este fin. Así, se exponen los métodos para determinar la validez y la confiabilidad y las normas de puntuación para los instrumentos y escalas de medición, subrayando sus ventajas, inconvenientes y factores condicionantes para su uso, así como las fuentes básicas de consulta para su correcta aplicación.

Palabras clave: Escalas; Construcción de escalas; Confiabilidad; Validez; Psicología de la salud.

ABSTRACT

Although there are numerous available scales for psychological measurement, the differences among populations as well as the diversity of issues studied in health psychology, frequently requires constructing new scales. Unfortunately, the "rules of thumb" and traditional procedures used are, in many cases, inadequate. Several such procedures are ruled out by recent psychometrical evidence. In this review, an updated state of the art on scale construction procedures for psychological instruments is presented in the context of the nature of psychological measurement and the way it limits measurement procedures and scale construction. The present review includes methods to determine validity, reliability and standardized scores, with emphasis on their advantages, inconveniences, conditioning factors and basic references for their optimum use.

Key words: Scales; Construction of scales; Validity; Reliability; Health psychology.

Al igual que el resto de las ciencias, la psicología requiere medir los fenómenos que estudia; sin embargo, a diferencia de lo que ocurre en las ciencias físicas y naturales, la medición en psicología se topa con dos problemas fundamentales: el primero es que muchos fenómenos psicológicos no son observables (aunque pueden serlo de manera indirecta mediante algunas de sus manifestaciones), y el

[1] Instituto de Investigaciones Psicológicas, Universidad Veracruzana, Edif. C, 2° piso, Av. Dr. Luís Castelazo Ayala s/n, Col. Industrial Ánimas, 91190 Xalapa, Ver., México, tel. (228)841-89-00, correo electrónico: rlc.academico@yahoo.com.mx. Artículo recibido el 26 de octubre y aceptado el 15 de diciembre de 2015.

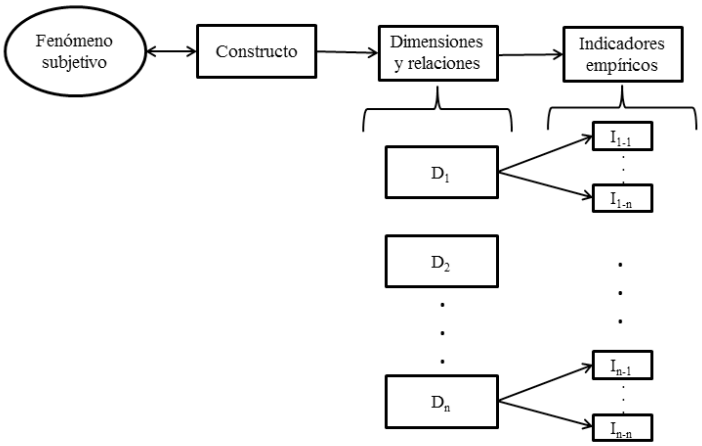
segundo es que es muy difícil establecer unidades de medición para dichos fenómenos. Para llevar a cabo mediciones en psicología es preciso resolver ambos problemas.

Aunque los fenómenos psicológicos subjetivos no sean observables, es posible plantear situaciones reales o simuladas en las que estos se manifiesten de manera más o menos objetiva. Bajo este supuesto, el objetivo planteado en esta revisión es presentar un panorama actualizado de dichos procedimientos, empezando por una breve discusión de la naturaleza de la medición en la psicología y la manera en que la misma condiciona la medición de las variables psicológicas y la construcción de instrumentos para este fin. A continuación, se presenta una exposición de los métodos para determinar la validez, la confiabilidad y las normas de puntuación para instrumentos y escalas de medición, enfatizando sus ventajas, inconvenientes y factores condicionantes.

En términos generales, la medición de los referidos fenómenos se lleva a cabo mediante un proceso que involucra dos pasos: *1)* la creación de un concepto o constructo riguroso que defina claramente el fenómeno que se quiere medir, el cual debe plantear las definiciones, dimensiones, conceptos subsidiarios e interrelaciones del fenómeno de la manera más detallada posible, y *2)* con base en esos constructos, se plantean situaciones observables en las cuales la manifestación del fenómeno de interés afecta la conducta de los individuos (Hernández, Fernández y Baptista, 2010).

En la psicología, con frecuencia se tiene la posibilidad de plantear situaciones supuestas en el papel o de manera oral para preguntar a los individuos por su grado de acuerdo con ellas o si es que se ajustan o describen su situación personal con el tema investigado. Sus respuestas permiten obtener una medida del constructo. En general, se asume que si se crea el clima de confianza adecuado con la persona que responde las preguntas, si se le garantiza la confidencialidad de sus respuestas y se le solicita su cooperación sincera y sin prisas, las respuestas que dará a las preguntas serán un reflejo fiel del proceso subjetivo, siempre que el cuestionario o instrumento a contestar sea válido y confiable. El proceso general se resume en la Figura 1.

Figura 1. El proceso de obtener indicadores empíricos para medir un fenómeno subjetivo. En primer término, es necesaria una caracterización del fenómeno a través del constructo teórico. A su vez, ese constructo está integrado por una serie de dimensiones relacionadas entre sí. Es a partir de tales dimensiones y relaciones que es posible crear los reactivos para reflejarlos en las situaciones en que el fenómeno se manifiesta.



Obsérvese que en la figura se plantea una relación recíproca entre el fenómeno subjetivo y el constructo, porque el constructo mismo se robustece y modifica con el estudio del fenómeno, y a la vez indica las relaciones fundamentales que se conocen sobre el fenómeno a estudiar y condiciona la manera de aproximarse a él.

En la sección correspondiente a la validez se abordarán dos cuestiones fundamentales relacionadas con el esquema planteado: *1)* la comproba-

ción de que los reactivos construidos representan realmente a todas las dimensiones del fenómeno, y *2)* si el instrumento resultante reproduce adecuadamente la estructura y relaciones del constructo en el que está basado.

LA UNIDAD DE MEDIDA DE LOS FENÓMENOS PSICOLÓGICOS

Aunque sea posible caracterizar y medir un fenómeno psicológico, persiste la duda sobre cuál es su unidad de medida. El problema es complicado porque no hay una manera unívoca de definir los fenómenos psicológicos. Por el contrario, hay una multitud de teorías y definiciones sobre lo que puede ser la inteligencia o la personalidad, por poner dos ejemplos, y cada una considera dimensiones y relaciones que pueden no aparecer en la otra. La situación se complica todavía más porque a partir de una definición es posible construir muchos reactivos e instrumentos diferentes, cada uno con su propia escala de puntuación.

Tal multiplicidad de definiciones implica que tampoco hay un cero absoluto conocido para cualquier fenómeno psicológico. La falta de un cero absoluto impide la construcción de una unidad consensuada porque no hay un punto de partida para considerar la ausencia del fenómeno psicológico. Aunque un individuo no conteste ninguna de las afirmaciones o preguntas de una escala o instrumento, no es posible deducir de ello que carece del rasgo que se pretende medir.

Para lidiar con estas dificultades, se han construido dos teorías generales que sustentan la construcción de instrumentos en psicología: la teoría clásica de los tests (TCT) y la teoría de respuesta al ítem (TRI) (Muñiz, 2010). La descripción detallada de dichas teorías sobrepasa los objetivos de este artículo, pero se considerará brevemente cómo es que ambas resuelven el problema de las unidades de medida de los fenómenos psicológicos.

El planteamiento fundamental de la TCT es que la medición de un fenómeno psicológico *es una función de cada instrumento considerado como un todo*; es decir, cada uno de los reactivos del instrumento, junto con los demás, aporta a la medición del constructo, y si se modifican uno o varios reactivos del mismo, se tiene un instrumento de medición con propiedades diferentes (Muñiz, 2010). Ello tiene como consecuencia que la medida del fenómeno dependa del instrumento utilizado. Cada instrumento tiene diversas fuentes de error en la medida, y no es posible comparar las mediciones obtenidas con dos instrumentos diferentes, aunque se apliquen a la misma persona y en condiciones idénticas. Por otro lado, los atributos fundamentales de los instrumentos (su confiabilidad, validez y normas de calificación) construidos con los procedimientos de la teoría clásica no son propiedades intrínsecas de los mismos sino que dependen del grupo normativo del que se obtienen. Por ejemplo, una escala que mida los conocimientos sobre algún tema será muy sencilla si la contestan individuos expertos, y muy difícil si la responden aquellos que lo desconocen. Dichas situaciones inciden directamente en las normas de calificación y en las interpretaciones de los puntajes que se obtienen con ellas (Prieto y Delgado, 2010).

Para evitar estos problemas, la TCT recomienda nunca comparar las calificaciones brutas obtenidas con diferentes instrumentos, y utilizar las normas de calificación solamente para individuos con características muy similares a las del grupo en el que se definieron. En la práctica, esto significa que hay que recalcular normas e indicadores de confiabilidad y validez con cada grupo en el que se aplique el instrumento, lo cual es psicométrica y pragmáticamente sensato, pero no deja de ser insatisfactorio desde el punto de vista científico (Muñiz, 1997). Por otra parte, hay trabajos (Henson y Roberts, 2006; Juárez, Idrovo, Camacho y Placencia, 2014) que muestran que una parte considerable de los trabajos de investigación en la psicología de la salud no proporcionan dato alguno acerca de su confiabilidad y validez; lo cual puede condicionar (o en casos extremos invalidar) las conclusiones obtenidas en esas investigaciones.

La TRI, en cambio, plantea que la medición del fenómeno psicológico es una función de cada reactivo por separado, y que cada uno de ellos tiene una función de información y una curva característica que permite detectar el nivel del rasgo en función de la respuesta dada por el individuo. Esta teoría es la más utilizada en la elaboración de pruebas de ejecución máxima (Muñiz, 2010) porque permite construir instrumentos que son independientes de las características del cuestionario específico utili-

zado y del grupo al que se le aplica. Pese a ello, su uso requiere que el constructo a medir y las condiciones de la aplicación satisfagan criterios muy restrictivos que no son fáciles de obtener en la práctica: que el rasgo a medir sea unidimensional y que se cumpla con la independencia local de los reactivos (es decir, que las respuestas que se dan a cada reactivo no se vean influidas por las que se dan a los otros). Estas restricciones han hecho a la TRI poco adecuada para construir instrumentos que midan constructos multidimensionales como los que suelen analizarse en la psicología de la salud (Muñiz, 1997). Sin embargo, se están probando programas para trabajar constructos y reactivos multidimensionales con la TRI, lo cual podría impulsar la investigación en este campo en un futuro cercano (Chalmers, 2012; Han y Paek, 2014).

CARACTERÍSTICAS FUNDAMENTALES DE LOS INSTRUMENTOS DE MEDICIÓN

Una vez que se ha obtenido un cuestionario que permite obtener una medida de un rasgo o de un fenómeno psicológico, debe contestarse tres preguntas básicas sobre la medida obtenida: ¿El instrumento mide realmente lo que tiene que medir? ¿La medida del instrumento es estable y cómo la afectan factores aleatorios? ¿Dónde se ubica la puntuación de un individuo particular en relación con los demás?

La primera pregunta corresponde a la validez del instrumento, la segunda a la confiabilidad y la tercera a las normas de puntuación.

Validez

La definición tradicional de la validez de un instrumento de medida es que el instrumento mida en efecto lo que tiene que medir. Sin embargo, la definición actual de validez incluye gran cantidad de aspectos y es mucho más concreta: "El grado en el cual las conclusiones e interpretaciones de cualquier medida están bien conformadas y justificadas, en tanto que son a la vez significativas y relevantes" (Cook y Beckman, 2006). Para conocer realmente la validez de un instrumento, se debe hacer un juicio evaluativo global sobre si los datos empíricos y los constructos respaldan la pertinencia y el significado de las interpretaciones que se hacen con base en los puntajes de las pruebas (Oluwatayo, 2012). Por otra parte, la validez de un instrumento no se establece de una vez por todas, sino que es resultado del acopio de evidencias y constructos que se logra en un proceso continuo (Aliaga, 2006).

Se ha argumentado que la validez es una propiedad unitaria, referida a las interpretaciones y usos que se hacen de las puntuaciones obtenidas al aplicar un instrumento (Messick, 1991). Pero para entender sus diferentes aspectos y la manera en que se evalúan es pertinente exponerlos por separado. La exposición que se hace a continuación está basada en el trabajo de Batista, Coenders y Alonso (2004). En cada apartado se describen los tipos de validez y una discusión breve de los procedimientos actuales para evaluarlos.

Validez de facie

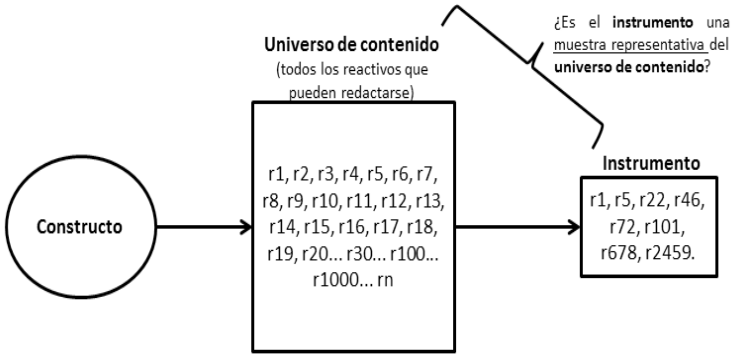
La validez *de facie* (llamada también de aspecto o aparente) se basa en juicios subjetivos –tanto del constructor como de los usuarios– sobre si el instrumento verdaderamente *parece* una encuesta formal o un instrumento de medida. Esto incluye un juicio valorativo sobre el léxico empleado, la claridad de las instrucciones, la organización del instrumento y la consideración sobre si sus reactivos son relevantes, claros, entendibles y razonables (Oluwatayo, 2012). La falta de validez *de facie* produce críticas y resistencia por parte de las personas que contestan el instrumento. Oluwatayo (2012) proporciona una descripción de los principales criterios que conviene cuidar para que el instrumento tenga validez *de facie*: *1)* que el formato sea claro y congruente con la estructura de un instrumento genuino; *2)* que los reactivos sean claros, sin ambigüedad, y de un nivel de dificultad apropiado para quien responde; *3)* que el deletreo de los términos difíciles y el espaciado de líneas sean los correctos; *4)* que las instrucciones sean claras, suficientes y adecuadas; *5)* que los reactivos parezcan razonables de acuerdo con el propósito del instrumento, y *6)* que la impresión sea clara y la calidad del papel la adecuada.

Validez de contenido

Esta característica responde a la pregunta de si los reactivos incluidos representan realmente todas

las dimensiones del fenómeno. Su análisis determina qué tan adecuado es el muestreo que hace un cuestionario del universo de posibles conductas que reflejan el constructo, y previene asimismo uno de los principales problemas que aparecen cuando solamente se cuida la validez de constructo: la subrepresentación del dominio del instrumento (Cohen y Swerdik, 2001; Messick, 1991). Esto ocurre porque los reactivos de todo instrumento, por muy extenso que sea, constituyen solo una muestra de todos los reactivos que podrían construirse (universo de contenido) (Figura 2).

Figura 2. ¿Los reactivos que constituyen el instrumento son verdaderamente representativos del universo de contenido? A partir del constructo elaborado, se puede construir una gran cantidad de reactivos que reflejen todas las dimensiones y relaciones posibles. Una muestra de reactivos que no sea representativa dejará sin cubrir aspectos fundamentales del constructo, y afecta directamente la validez global del instrumento.



Dado lo anterior, es necesario asegurarse de que la muestra de reactivos que constituyen el cuestionario sea representativa e incluya todas las dimensiones del constructo. Desafortunadamente, no hay una manera directa de hacer lo anterior. La validez de contenido se estima por métodos indirectos, y el más común y eficiente de ellos es el *juicio de expertos* (Escobar y Cuervo, 2008; Muñiz, 1997).

El juicio de expertos involucra la revisión del instrumento por parte de especialistas con experiencia en el trabajo con el constructo de interés, quienes juzgarán si los reactivos son una muestra equilibrada y representativa del universo de contenido. En su revisión de 2008, Escobar y Cuervo reseñan las cualidades que los jueces elegidos deberán tener para realizar un trabajo óptimo: *1)* experiencia en la realización de juicios y toma de decisiones basadas en evidencias, *2)* reputación en la comunidad científica, *3)* disposición y motivación para participar, y *4)* imparcialidad, confianza en sí mismos y adaptabilidad. Además, las autoras recomiendan que al menos uno de los jueces sea un lingüista.

Los aspectos a evaluar por los jueces pueden variar, al igual que los criterios para hacerlo. Escobar y Cuervo (2008) señalan la necesidad de evaluar al menos si cada uno de los reactivos pertenece verdaderamente al constructo y a la dimensión para la cual se construyó, la tendenciosidad o sesgo del reactivo y la adecuación lingüística del reactivo para la población a la que va destinado. Pero existen otros métodos, como el de Lawshe (1975), que recomienda que los jueces califiquen si un reactivo es necesario para el objetivo del cuestionario, si es útil pero no necesario, o si carece de utilidad.

La integración de los juicios para obtener un consenso (conocida también como concordancia entre jueces) tampoco tiene una solución única. Las formas más aceptadas de solucionar el problema involucran medidas de concordancia, como el coeficiente Kappa para variables dicotómicas (Cohen, 1960), las extensiones del mismo para variables politómicas, el coeficiente de correlación intraclase para variables continuas (Shoukri, 2004), el coeficiente de validez de contenido (Lawshe, 1975) o las medidas loglineares, como el enfoque *mixture* (Ato, Benavente y López, 2006)[2].

Siempre debe tenerse en cuenta que el objetivo final del proceso de jueceo es obtener un ins-

[2] Para una revisión detallada de las medidas de concordancia, véase Gwet (2012). Los pasos para realizar un análisis de validez de contenido vienen adecuadamente descritos en los trabajos de Haynes, Richard y Kubany (1995) y de Escobar y Cuervo (2008).

trumento cuyos reactivos representen adecuadamente el constructo, y que esté redactado con un lenguaje comprensible para el grupo de población al que se va a aplicar y con un mínimo de sesgo inducido.

Validez de constructo

Este tipo de validez se considera fundamental para la evaluación de escalas e instrumentos, ya que los análisis para determinarla comprobarán si la estructura del instrumento reproduce realmente la del constructo planteado. Recuérdese que *el constructo planteado es una concreción de un fenómeno psicológico inobservable y que sus dimensiones y relaciones determinan todo aquello que es posible observar y medir del fenómeno*. La validez de constructo suele dividirse en validez nomológica, validez convergente y validez discriminante. La validez *nomológica* se refiere a que hay correspondencia entre las relaciones teóricas y las encontradas con el instrumento. Si la teoría, por ejemplo, plantea la existencia de tres subescalas o dimensiones, se esperaría que las puntuaciones de quienes respondan el instrumento se agrupen en tres subescalas. La validez *convergente* se refiere a que las medidas aportadas por el instrumento deben tener una correlación directa con las de otros instrumentos que miden el mismo constructo. Hay validez *discriminante*, por el contrario, si las medidas aportadas por el instrumento no tienen relación con las de otros instrumentos que evalúan otros constructos (Batista et al., 2004).

Los métodos más comunes para determinar la validez de constructo son el análisis factorial (en sus modalidades exploratoria y confirmatoria) y los coeficientes de correlación. El análisis factorial es una técnica estadística que permite determinar una cantidad reducida de factores comunes que expliquen la variabilidad de los datos (Beaver et al., 2013). Esta característica lo hace especialmente pertinente para determinar si la estructura del instrumento reproduce la de la teoría o el constructo que se utilizó para elaborarlo. Los datos necesarios son las puntuaciones de los reactivos, *y se espera que los factores obtenidos correspondan con las dimensiones del constructo planteado*.

A continuación se describen someramente sus aspectos principales y las recomendaciones generales para su uso e interpretación apropiados[3].

a) *Análisis factorial exploratorio* (AFE). El AFE es una técnica paramétrica y su uso requiere que los datos se comporten linealmente y sigan una distribución normal multivariante (Ferrando y Anguiano, 2010). Es frecuente que se revise el supuesto de normalidad mediante la aplicación de una prueba de bondad de ajuste –como la de Kolmogorov-Smirnov– y el de linealidad a través del examen de los cocientes de asimetría de los reactivos (Ferrando y Anguiano, 2010). Cuando no se cumple con el supuesto de normalidad, se pueden utilizar versiones del análisis que emplean métodos de extracción de factores que permiten obviarlo. El segundo requisito es más delicado, pues una asimetría muy marcada y bidireccional en los reactivos puede requerir el uso de técnicas de análisis no lineales. El análisis factorial funciona bien si todos los reactivos tienen cocientes de asimetría en el intervalo de +1 a –1, lo cual es más sencillo cuando se utilizan reactivos de respuesta graduada, como los de tipo Likert (Ferrando y Anguiano, 2010).

El segundo paso es confirmar si la muestra de aplicaciones obtenida es adecuada para llevar a cabo la técnica, lo que se realiza mediante las pruebas de Kaiser-Meyer-Olkin (KMO), la prueba de esfericidad de Bartlett y el cálculo de la determinante de la matriz de correlaciones. Es necesario saber, en primer lugar, si las variables están correlacionadas de manera significativa o no. La prueba de esfericidad de Bartlett permite conocer lo anterior al contrastar la hipótesis nula de que los reactivos no correlacionan entre sí (esto es, que la matriz de correlaciones es una matriz singular). Si no se puede rechazar la hipótesis nula, los factores obtenidos por el análisis serán completamente espurios. Si se rechaza la hipótesis nula, la determinante de la matriz de correlaciones indicará si es posible extraer un número limitado de factores (que representarán las dimensiones del constructo). Si el valor de la determinante es cero o negativo, no es posible realizar la extracción y no puede continuarse con el análisis factorial. Por último, es ne-

[3] Al respecto, el lector interesado puede consultar los trabajos de Ferrando y Anguiano (2010), Beaver et al. (2013), Gaskin y Happell (2014) y Lloret, Hernández y Tomás (2014).

cesario conocer el grado de correlación conjunta que tienen las variables, lo que se hace mediante la prueba KMO. Los valores óptimos abarcan de 0.8 a 1. Se estima que un valor de 0.65 a 0.8 puede ser útil, pero condiciona la interpretación de los resultados (Beaver et al., 2013).

El tercer paso, esto es, la extracción de los factores, es de importancia crítica. Es necesario recordar que los factores se corresponderán con las dimensiones del constructo de base empleado. Desafortunadamente, este proceso es complejo, por lo cual se presta a la comisión de muchos errores, de los cuales el más repetido y extendido parece ser el uso del análisis de componentes principales (ACP) como método de extracción de factores (Gaskin y Happell, 2014; Henson y Roberts, 2006; Juárez et al., 2014). El ACP emplea la varianza total de los reactivos, que incluye entre sus cálculos la varianza de error; en cambio, los procedimientos de análisis factorial solamente utilizan la varianza común para explicar los datos, por lo que son mucho más convenientes para hacer la extracción de factores en cuanto que previenen la posibilidad de obtener factores espurios y varianzas explicadas infladas, como se ha confirmado en estudios de simulación (Beaver et al., 2013; Ferrando y Anguiano, 2010). Cuando se cumple el requisito de normalidad multivariante, los métodos de extracción más utilizados son el de máxima verosimilitud y el de mínimos cuadrados generalizados. Cuando se incumple ese supuesto, los métodos más adecuados son el de mínimos cuadrados no ponderados y el de factorización de ejes principales (Briggs y McCallum, 2003). Además, el método de mínimos cuadrados no ponderados tiene la gran ventaja de rescatar factores débiles y generar buenos resultados incluso con muestras muy pequeñas, lo que es especialmente útil al trabajar con poblaciones muy específicas (de entre cincuenta y cien individuos, y hasta menos de treinta si el número de factores es pequeño) (Jung, 2013; Ximénez y García, 2005).

Como resultado de la extracción de factores se obtendrán dos tablas: una donde se detallará el número de factores obtenidos, con la varianza explicada por cada uno, y otra donde se especificará la estructura general del cuestionario, con las cargas factoriales de todos los reactivos y el grado en que correlacionan con cada factor (véase un ejemplo del segundo tipo de tabla en el Cuadro 1).

Cuadro 1. Ejemplo de tabla de cargas factoriales del análisis factorial exploratorio. Nótese que los primeros seis reactivos presentan una carga factorial muy clara hacia uno de los factores. El reactivo 7 presenta ambigüedad y debería ser eliminado de la versión final del instrumento, a menos que existan razones teóricas o empíricas importantes para retenerlo.

Reactivo	Factor 1	Factor 2
1	0.835	0.023
2	0.667	0.250
3	0.711	0.172
4	0.076	0.916
5	0.356	0.627
6	0.226	0.701
7	0.482	0.503

Debido a las operaciones matemáticas necesarias para realizar el análisis, en las soluciones extraídas en primera instancia las varianzas explicadas por los factores suelen estar muy cargadas hacia el primer factor. Para eliminar este sesgo, es necesario recurrir al procedimiento de rotación de los factores.

Existen varios métodos de rotación, los que se pueden agrupar en ortogonales y oblicuos. Los métodos ortogonales asumen que los factores obtenidos no están correlacionados entre sí, esto es, que son independientes; los oblicuos asumen que existe un grado de correlación apreciable entre los factores y ayudan mucho a clarificar la estructura del instrumento cuando este es efectivamente el caso. Entre los métodos ortogonales, el más utilizado es el Varimax, pues se considera sólido, estable y sencillo de interpretar (Fabrigar, Wegener, MacCallum y Strahan, 1999). Entre los métodos oblicuos están el Oblimin Directo y el Promax. Ambos métodos son aceptables, pudiéndose utilizar los valores que los programas comerciales incorporan por defecto. En la psicología, muchos de los constructos que se estudian plantean que sus dimensiones se relacionan entre sí; por ello, resulta una buena idea iniciar el análisis utilizando un método de rotación oblicua, y si las correlaciones entre los factores son inferiores a 0.32, se puede realizar

una nueva extracción utilizando un método ortogonal (Beaver et al., 2013; Ferrando y Anguiano, 2010).

Para obtener una solución final es necesario saber cuántos factores hay que retener. En la literatura relacionada se detallan varios criterios, entre los cuales destacan la regla de Kaiser y el gráfico de sedimentación (también llamado de guijarros o *scree-plot*), que se basan en los autovalores (*eigenvalue*) de los factores. Tanto la regla de Kaiser como el gráfico de sedimentación tienen el mismo problema: se basan en la lógica del ACP y utilizan la varianza total y no la varianza común para calcular los autovalores de los factores. Cuando se emplea el análisis factorial propiamente dicho, los autovalores carecen de sentido y no pueden interpretarse de manera adecuada (Ferrando y Anguiano, 2010; Lorenzo, Timmerman y Kiers, 2011). En su lugar, se ha propuesto el uso del análisis paralelo, que consiste en comparar los autovalores de los factores obtenidos con los de una muestra de autovalores generados al azar (Horn, 1965). Aunque el uso de esta técnica ha estado sujeta a críticas (Lorenzo et al., 2011; Ruscio y Roche, 2012), existen desarrollos promisorios recientes que la hacen más robusta y que contribuyen a eliminar su defecto principal: sobreestimar el número de factores a retener (Green, Levy, Thompson, Lu y Wen-Juo, 2012; Green, Thompson, Levy y Wen-Juo, 2015). Desafortunadamente, esta técnica no está disponible en la mayor parte de los programas estadísticos comerciales, pero Hayton, Allen y Scarpello (2004) proponen un sencillo tutorial para incorporarlo como una rutina en el lenguaje de SPSS.

En la práctica, los métodos más utilizados son el uso de las cargas factoriales de los reactivos, que pone una atención cuidadosa al contenido de los mismos, y la teoría o constructo de base, cuando existe (Ferrando y Anguiano, 2010; Lloret et al., 2014). Para conservar un factor, debería tener al menos a tres reactivos con cargas factoriales superiores a 0.35 (Ferrando y Anguiano, 2010). Es prudente también eliminar los reactivos con carga ambigua (cargas superiores a 0.35 en un factor y a 0.25 en otro [véase el reactivo 7 del Cuadro 1]). Los factores con dos o menos reactivos pueden eliminarse, a menos que se tengan buenas razones para conservarlos (puede ser el caso de factores heurísticamente valiosos o con índices de confiabilidad elevados). El proceso de eliminación y extracción se debe repetir cuantas veces sea necesario para obtener una solución clara, fácilmente interpretable y con el mínimo posible de reactivos ambiguos. Por supuesto, se espera que la estructura encontrada reproduzca la de la teoría o constructo de base.

b) Análisis factorial confirmatorio (AFC)

La variedad confirmatoria del análisis factorial puede considerarse como una auténtica prueba de hipótesis. En el dominio de la construcción de instrumentos, se utiliza para contrastar la hipótesis de que un instrumento particular se adapta a una estructura determinada por la teoría o las observaciones (Lloret et al., 2014). Esta técnica es especialmente valiosa cuando se cuenta con un constructo teórico de base muy robusta, cuando se desea comprobar la validez de una adaptación transcultural, o cuando se tienen dos o más constructos "en conflicto", y se quiere saber cuál es el que se ajusta mejor a los datos empíricos proporcionados por el instrumento. El AFC es un procedimiento complejo cuya descripción detallada excede los alcances de este artículo, por lo cual nos limitaremos a ofrecer las principales directrices para su utilización[4].

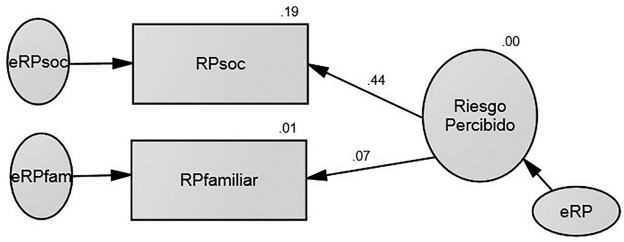
Planteamiento del constructo de base. Se refiere a la especificación de un constructo o teoría de base que se comparará con los datos aportados por el cuestionario. Si se está en el proceso de construir uno, se tendrá resuelto este paso en las primeras etapas.

Especificación del modelo. Este paso dependerá del software utilizado para hacer los análisis. Algunos programas permiten hacer esos análisis dibujando simplemente el diagrama del constructo de base, lo que se ejemplifica de manera muy simplificada en la Figura 3. En ella se muestran los símbolos utilizados.

Los rectángulos son respuestas observables (indicadores empíricos o puntajes de reactivos); los círculos u óvalos son variables inobservables (que pueden ser dimensiones, constructos o términos de error); las flechas unidireccionales plantean rela-

[4] El lector interesado puede consultar trabajos como el de Schreiber, Stage, King, Nora y Barlow (2006), o el manual de Arbuckle (2010). Aquí se enlistan solamente los pasos principales.

Figura 3. Ejemplo de análisis factorial confirmatorio. El constructo (muy simplificado) está representado por dos reactivos, que son las únicas puntuaciones empíricas que se obtienen de los sujetos (rectángulos, respuestas observables). Ambas respuestas son medidas indirectas del constructo inobservable "riesgo percibido" (óvalo). Los términos de error (eRPsoc, eRPfam y eRP) son también inobservables y, al igual que el constructo principal, determinan la puntuación de los sujetos en ambos reactivos. Las flechas dan cuenta de esta relación de determinación unidireccional. Para mayores detalles véase el texto.



ciones directas y no recíprocas entre variables y, aunque no se muestran en el ejemplo, las flechas bidireccionales se utilizan para indicar correlación entre dos variables.

Obtención de las relaciones entre las variables. Tras el paso anterior se puede proceder al análisis, que dará generalmente dos tipos de resultados: los coeficientes de correlación entre las diferentes variables, y la varianza explicada por las variables inobservables que convergen en una variable observable en particular. En la Figura 3 se presenta un ejemplo de lo anterior en una escala de solamente dos reactivos. Encima de cada línea se muestra el coeficiente de correlación correspondiente, y encima de cada cuadro la varianza explicada por la variable con la que está relacionado. El factor común o constructo de base explica mucho más varianza de la variable RPsoc (19%) que de RPfamiliar (apenas 1%).

Ajuste de los datos al modelo propuesto. Este es el verdadero paso crítico del AFC, pues proporciona los índices que contrastan la hipótesis de que los datos del cuestionario se ajustan al modelo teórico especificado. La mayoría de los programas proporciona un gran número de índices de ajuste. No existe un índice de ajuste perfecto ni completamente consensuado, pero se ha demostrado en estudios de simulación que el índice comparativo de ajuste (*comparative fit index*, o CFI) y la raíz cuadrada del error de aproximación (*root mean square error of aproximation*, o RMSEA) predicen, en conjunto, un adecuado ajuste de los datos al modelo. Los valores óptimos de ambos índices son, a saber: CFI > 0.90; RMSEA ≥ 0.03 y ≤ 0.08 (Schreiber et al., 2006).

Una de las grandes ventajas del AFC es que permite decidir cuál de los distintos modelos teóricos en conflicto se ajusta mejor a una serie de datos empíricos (Lloret et al., 2014). Así, el investigador puede, por ejemplo, decidir si su constructo está representado de manera óptima por dos, tres o más factores en la población que estudia. Esto es particularmente útil cuando en la literatura se reportan varias estructuras factoriales diferentes y el investigador debe de decidir cuál se ajusta mejor a sus datos. Esto se hace especificando cada uno de los modelos en el programa y comparando sus índices de ajuste respectivos. El modelo con mejores índices de ajuste será el que mejor explique los datos. Una nota precautoria: siempre debe tenerse en cuenta que con muestras grandes y reactivos ambiguos (que tienen carga en más de un factor), el AFC puede mostrar índices de ajuste inadecuados, aun cuando no lo sean realmente (Lloret et al., 2014). Ya Ferrando y Lorenzo (2000) han propuesto algunos procedimientos para lidiar con este problema.

Modificaciones al modelo propuesto. Cuando los índices de ajuste no son adecuados, conviene eliminar los reactivos (e incluso las dimensiones) con menores correlaciones o varianza explicada. Una poda cuidadosa de los reactivos puede mejorar los índices de ajuste, pero debe tenerse en cuenta que cuando la muestra es demasiado grande, cualquier mínima discrepancia entre los datos y el constructo dará lugar a índices inaceptables (Lloret et al.,

2014). Esto puede solucionarse dividiendo la muestra en partes iguales y haciendo el procedimiento en cada una, o bien tomando al azar una muestra más pequeña de los datos recopilados.

Validez predictiva

Este tipo de validez hace referencia a que los resultados de la aplicación del instrumento concuerdan con los que se obtienen de estudios empíricos. En términos científicos, tiene una importancia crítica porque demuestra que las mediciones no únicamente reproducen la teoría, sino también los resultados empíricos obtenidos con otro tipo de estudios. Por ejemplo, si la teoría predice que los puntajes del rasgo medido diferirán entre hombres y mujeres, los puntajes obtenidos con el instrumento *deben* de ser diferentes para hombres y mujeres. La validez predictiva de un instrumento aumenta a medida que aumentan los resultados que avalan su desempeño satisfactorio o su capacidad de anticipar una situación o resultado (Oluwatayo, 2012).

Confiabilidad

Un cuestionario es confiable si la medida que proporciona es estable a través del tiempo, o si dos o más evaluadores obtienen la misma medida en el mismo momento de aplicación. La confiabilidad es muy útil para determinar qué tan precisa es la medición obtenida con el instrumento (Houser, 2008). Como la confiabilidad del instrumento puede verse afectada por una gran diversidad de factores, las puntuaciones de los individuos siempre están sujetas a errores aleatorios. Nunca se obtiene la puntuación *verdadera* de la característica que se pretende medir. Debido a ello, no existe un procedimiento para obtener directamente la confiabilidad de una escala o instrumento de medición; pero sí puede estimarse por medio de distintos procedimientos estadísticos, todos los cuales se basan en el uso de coeficientes de correlación (Aiken, 2003; Aliaga, 2006). La confiabilidad se reporta en forma de un coeficiente que varía entre 0 (ausencia total de confiabilidad) y 1 (repetitividad perfecta de la medición).

En términos psicométricos, el mayor valor del coeficiente de confiabilidad es que permite *estimar* la puntuación verdadera de las personas *para un nivel de confianza dado*. Para lograrlo, se debe calcular primero el coeficiente de confiabilidad mediante alguno de los procedimientos que se reseñan más adelante. Luego, debe calcularse el error estándar de medición (EEM) utilizando la desviación estándar de las puntuaciones obtenidas por los individuos (s), y el coeficiente de confiabilidad calculado (rxx) por medio de la fórmula $EEM = s\sqrt{1 - rxx}$

Una vez obtenido el EEM, la puntuación verdadera (PV) de cada individuo se estima de la puntuación obtenida en la prueba (PO) para un nivel de confianza dado por medio de las fórmulas:

$PV = PO \pm EEM$ (Con 68% de confianza).

$PV = PO \pm 2\ EEM$ (Con 95% de confianza).

$PV = PO \pm 3\ EEM$ (Con 99% de confianza).

Observando lo anterior, es fácil entender la importancia de que un instrumento tenga un coeficiente de confiabilidad alto. Mientras mayor sea su confiabilidad, mayor será la precisión con la que estime la puntuación verdadera del individuo para un nivel de confianza dado. En la misma proporción, serán más precisas las decisiones que se tomen con las puntuaciones del instrumento, en el caso de que sea válido. Considérese como ejemplo un instrumento de percepción de riesgo con reactivos tipo Likert y puntuaciones totales con valores de entre 10 y 50, el cual se aplica a una muestra de trescientas personas. Tras la aplicación, se obtiene una media de las puntuaciones de 25 y una desviación estándar de 5. Como también interesa clasificar a los individuos en niveles de percepción, se construyen tres categorías basadas en la división en terciles de las puntuaciones obtenidas; así, una persona con puntación por arriba de 33 tiene un puntaje de percepción de riesgo alto, otra con puntaje de 21 a 32 un puntaje intermedio, y 20 o menos corresponde a una baja percepción de riesgo. Imaginemos ahora a un individuo con una puntación obtenida de 27, el cual calificaría en el grupo de percepción de riesgo intermedio. Por lo anteriormente discutido, se sabe que esta puntuación no corresponde a su puntuación real, pero la misma puede ser estimada por medio de las fórmulas reseñadas. Si la escala tiene un coeficiente de confiabilidad de 0.90, la puntuación verdadera debe estar, con una probabilidad de 95%, en un rango de 27 ± 2 (1.5), es decir, entre 30 y 24. Cualquiera de estos puntajes sigue situando a esa persona en la categoría de riesgo intermedio. Pero si el instru-

mento tiene coeficiente de confiabilidad de 0.50, los valores anteriores se transforman a 27 ± 2 (3.5), y en consecuencia su puntuación verdadera estará, con un 95% de probabilidad, entre 34 y 20. Es decir, *la puntuación verdadera podría corresponder a cualquiera de las tres categorías*. Casi está de más de decir que para un instrumento tan poco confiable, la clasificación en categorías carece de sentido (y asimismo las decisiones que se tomen con base en ellas).

Los métodos más habituales para estimar la confiabilidad se reseñan a continuación[5].

Confiabilidad test-retest. Se aplica el mismo instrumento a los mismos individuos en dos momentos lo suficientemente separados en el tiempo como para que los efectos de la memorización se reduzcan al mínimo. Las puntuaciones obtenidas se correlacionan con la técnica apropiada (generalmente los coeficientes de correlación de Pearson o Spearman). Es el método más sólido para rasgos que son estables en el tiempo.

Formas paralelas. Las formas paralelas son instrumentos construidos para medir el mismo constructo con diferentes reactivos. Se considera que las dos o más formas paralelas son el mismo instrumento, lo que permite reducir considerablemente el intervalo que debe transcurrir entre las dos aplicaciones. Tiene el inconveniente de que es necesario demostrar la equivalencia en contenido y estructura de las formas que se utilicen, lo cual puede hacerse mediante técnicas de jueceo y análisis factorial confirmatorio.

División en mitades emparejadas. Este método consiste en dividir la prueba en dos mitades para determinar el coeficiente de confiabilidad entre las puntuaciones de ambas. Se basa en la lógica de que, como todos los reactivos están en la misma escala y miden el mismo constructo, se pueden obtener dos formas paralelas de la misma prueba. La división en mitades puede hacerse por el método de pares y nones (reactivos 1, 3, 5… vs. 2, 4, 6…) o con el primer 50% contra el segundo 50% de los reactivos de la prueba. Como la división acorta el instrumento, el coeficiente de correlación obtenido debe ajustarse por medio de la fórmula de Spearman-Brown (Aiken, 2003).

Método de la equivalencia racional (consistencia interna). Es una extensión del procedimiento anterior. Su planteamiento básico es que todos los reactivos pueden considerarse como instrumentos paralelos, y el coeficiente de correlación conjunto permitirá determinar la confiabilidad total de la prueba. La fórmula más general para determinar la confiabilidad por este procedimiento es el coeficiente alfa de Cronbach, que se ha convertido en una especie de "estándar de oro". Sin embargo, se aplica de manera incorrecta la mayor parte de las veces (Elosua y Zumbo, 2008). Como la fórmula para determinarlo emplea un coeficiente de correlación de producto-momento de Pearson, exige que los datos cumplan el requisito de normalidad multivariante para proporcionar resultados consistentes. En la práctica, tal supuesto se comprueba pocas veces, y existen trabajos que demuestran que el coeficiente alfa de Cronbach suele subestimar la confiabilidad cuando se aplica a datos que no lo satisfacen. En estos casos se deben emplear los otros procedimientos ya reseñados, o alguna de las formulaciones no paramétricas del citado coeficiente (Elosua y Zumbo, 2008).

Normas de puntuación

Las puntuaciones obtenidas en un instrumento generalmente requieren un proceso de interpretación o transformación. De acuerdo con la TCT, no es lícito utilizar las normas de puntuación obtenidas en un grupo para interpretar las puntuaciones de un individuo que pertenece a un grupo diferente debido a que las normas de puntuación son propiedades de la muestra en la que se han determinado, no del instrumento en sí mismo (Aiken, 2003). Para solucionar este problema se procede de dos formas: se transforman las puntuaciones a escalas normalizadas o se determinan puntos de corte apropiados para cada grupo en cada estudio que se realice (Muñiz, 1997).

Los métodos de transformación de los puntajes que más se utilizan consideran la conversión de las puntuaciones brutas obtenidas en rangos percentilares o en puntuaciones normalizadas Z o T. Para ello, se debe transformar el puntaje obtenido

[5] Una exposición detallada de los mismos puede consultarse en Anastasi y Urbina (1998) y Aiken (2003).

por cada persona utilizando los valores obtenidos dentro de su grupo de referencia[6]. Esta transformación permite, en teoría, comparar la puntuación de un individuo en particular con la de otro individuo de otro grupo de referencia, para lo cual se requiere administrar el instrumento a una muestra representativa de la población de la que proceda el sujeto con el fin de determinar los parámetros necesarios para la transformación.

En ocasiones no importa tanto comparar a dos individuos, sino establecer si la puntuación del cuestionario indica la presencia de un rasgo o una situación de interés, generalmente con propósitos clínicos o de diagnóstico. En este caso, los puntos de corte suelen determinarse con base en criterios externos, comparando las puntuaciones del instrumento con un "estándar de oro" con muy alta sensibilidad y especificidad, o bien con criterios clínicos ampliamente consensuados entre la comunidad de especialistas en el campo de interés. El punto de corte utilizado se determinará generalmente mediante el método de la curva ROC, tomando como tal aquel que maximice los valores de sensibilidad y especificidad (Cerda y Cifuentes, 2010). También en este caso será necesario determinar el punto de corte para cada población particular, toda vez que la prevalencia del rasgo o patología medidos afecta directamente los valores predictivos del punto de corte determinado (Jaeschke, Guyatt y Sackett, 1994). Un método alternativo, completamente referido a la muestra en la que se han determinado las puntuaciones, es utilizar la puntuación que corresponda al límite inferior del cuartil o tercil superior de la muestra recabada. Aunque no es un criterio objetivo, se utiliza de manera heurística en algunas investigaciones, siempre que no se comparen las puntuaciones brutas sino precisamente los límites inferiores de los cuartiles o terciles correspondientes (Aiken, 2003; Muñiz, 1997).

A modo de conclusión, debe resaltarse que pese a la existencia de una gran variedad de técnicas de recolección de datos en la investigación de la psicología de la salud, los instrumentos y escalas de medición en la psicología muestran múltiples ventajas, como su conveniencia; la facilidad de ser aplicados simultáneamente a una gran cantidad de personas; la facilidad para su calificación, clasificación e interpretación, y la multiplicidad de formatos que son posibles para su aplicación. Sin embargo, todas estas ventajas dependen de que su construcción, validación y uso se apeguen al empleo de técnicas psicométricas robustas debidamente contrastadas por estudios empíricos y con datos reales y simulados. Es posible considerar que las recomendaciones y la literatura reseñada en el presente trabajo contribuirán, como una referencia mínima, a cubrir estos aspectos de la construcción y la validación de escalas e instrumentos en la psicología de la salud.

REFERENCIAS

Aiken, L. (2003). *Tests psicológicos y evaluación* (11ª ed.). México: Pearson Educación.

Aliaga T., J. (2006). Psicometría: tests psicométricos, confiabilidad y validez. En A. Quintana y W. Montgomery (Eds.): *Psicología: Tópicos de actualidad*. Lima: UNMSM.

Anastasi, A. y Urbina, S. (1998). *Tests psicológicos* (7ª ed.). México: Prentice-Hall.

Arbuckle, J.L. (2010). *IBM SPSS® AmosTM 19 User's Guide.* Crewfordville, Fl: Amos Development Corporation.

Ato, M., Benavente, A. y López, J.J. (2006). Análisis comparativo de tres enfoques para evaluar el acuerdo entre observadores. *Psicothema*, *18*(3), 638-645.

Batista F., J.M., Coenders, G. y Alonso, J. (2004). Análisis factorial confirmatorio. Su utilidad en la validación de cuestionarios relacionados con la salud. *Medicina Clínica*. *122*(supl. 1), 21-27.

Beaver, A.S., Lounsbury, J.W., Richards, J.K., Huck, S.W., Skolits, G.J. y Esquivel, S.L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research & Evaluation*, *18*(6), 1-13.

Briggs, N.E. y MacCallum, R.C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research*, 38, 25-56.

Chalmers, R.P. (2012). MIRT: A multidimensional Item Response Theory Package for the R environment. *Journal of Statistical Software*, 48, 1-29.

[6] Para más información, consúltese Anastasi y Urbina (1998) y Aiken (2003).

Cerda J., L. y Cifuentes L., A. (2010). Uso de tests diagnósticos en la práctica clínica (Parte 1). Análisis de las propiedades de un test diagnóstico. *Revista Chilena de Infectología*, 27, 205-208.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cohen, R., Swerdlik, M. (2001). *Pruebas y evaluación psicológicas: Introducción a las pruebas y a la medición* (4ª ed.). México: McGraw-Hill.

Cook. D.A., Beckman, T.J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, 119, 166.e7-166.e16.

Escobar P., J. y Cuervo M., A. (2008). Validez de contenido y juicio de expertos: una aproximación a su utilización. *Avances en Medición*, 6, 27-36.

Elosua P., O. y Zumbo B., D. (2008). Coeficientes de fiabilidad para escalas de respuesta categórica ordenada. *Psicothema*, *20*(4), 896-901.

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C. y Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272-299.

Ferrando P., J. y Lorenzo S., U. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: some aspects of the problem and some suggestions. *Psicologica*, 21, 301-323.

Ferrando, P.J. y Anguiano C., C. (2010). El análisis factorial como técnica de investigación en Psicología. *Papeles del Psicólogo*, 31, 18-33.

Gaskin, C.J. y Happell, B. (2014). On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International Journal of Nursing Studies*, 51, 511-521.

Green, S.B., Levy, R., Thompson, M.S., Lu, M. y Wen-Juo, L. (2012). A proposed solution to the problem with using completely random data to assess the number of factors with parallel analysis. *Educational and Psychological Measurement*, 72, 377-393.

Green, S.B., Thompson, M.S., Levy, R. y Wen-Juo, L. (2015). Type I and type II error rates and overall accuracy of the revised parallel analysis method for determining the number of factors. *Educational and Psychological Measurement*, 75, 428-457.

Gwet, K.L. (2012). *Handbook of inter-rater reliability* (3th ed.). Gaithersburg, MD: Advanced Analytics, LLC.

Han, K.T. y Paek, I. (2014). A review of commercial software packages for multidimensional IRT modeling. *Applied Psychological Measurement* [On line], 1-13. doi: 10.1177/0146621614536770

Haynes, S.N., Richard, D.C.S. y Kubany, D.S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238-247.

Hayton, J.C., Allen, D.G. y Scarpello, V. (2004). Factor retention decisions in factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, 7, 191-205.

Henson, R.K. y Roberts, J.K. (2006). Use of factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, 66, 393-416.

Hernández S., R., Fernández C., C. y Baptista L., P. (2010). *Metodología de la investigación* (6ª ed.). México: McGraw-Hill.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 32, 179-185.

Houser, J. (2008). Precision, reliability and validity: essential elements of measurement in nursing research. *Journal for Specialists in Pediatric Nursing*, 13, 297-299.

Jaeschke, R., Guyatt, G. y Sackett, D.L. (1994). Users' guides to the medical literature: III. How to use an article about a diagnostic test: A. Are the results of the study valid? *Journal of the American Medical Association*, 271, 389-391.

Juárez G., A., Idrovo, A.J., Camacho Á., A. y Placencia R., O. (2014). Síndrome de burnout en población mexicana: una revisión sistemática. *Salud Mental*, 37, 159-176.

Jung, S. (2013). Exploratory factor analysis with small sample sizes: A comparison of three approaches. *Behavioural Processes*, 97, 90-95.

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28, 563-575.

Lorenzo S., U., Timmerman, M.E. y Kiers, H.A.L. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research*, 46, 340-364.

Lloret S., S., Ferreres T., A., Hernández B., A. y Tomás M., I. (2014). El análisis factorial exploratorio de los ítems: una guía práctica revisada y actualizada. *Anales de Psicología*, 30, 1151-1169.

Messick, S. (1991). Validity of test interpretation and use. En M.C. Alkin (Ed.): *Encyclopedia of Educational Research* (6th ed.). New York: McMillan.

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems*. Madrid: Pirámide.

Muñiz, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo*, *31*(1), 57-66.

Oluwatayo, J.A. (2012). Validity and reliability issues in educational research. *Journal of Educational and Social Research*, *2*(2), 391-400.

Prieto, G. y Delgado A., R. (2010). Fiabilidad y validez. *Papeles del Psicólogo*, *31*(1), 67-74.

Ruscio, J. y Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment*, 24, 282-292.

Schreiber, J.B., Stage, F.K., King, J., Nora, A. y Barlow, E.A. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. *The Journal of Educational Research*, *99*(6), 323-337.

Shoukri, M.M. (2004). *Measures of interobserver agreement*. Boca Raton, FLO: Chapman & Hall/CRC.

Ximénez M., C. y García A., G. (2005). Comparación de los métodos de estimación de máxima verosimilitud y mínimos cuadrados no ponderados en el análisis factorial confirmatorio mediante simulación Monte Carlo. *Psicothema*, 17, 528-535.