iMedPub Journals
http://imedpub.com

ARCHIVOS DE MEDICINA
ISSN 1698-9465

2015
Vol. 11 No. 3:1
**doi:** 10.3823/1251

# Construcción y validación de escalas de medición en salud: revisión de propiedades psicométricas

# Construction and validation of measurement scales in health: a review of psychometric properties

**Luján-Tangarife, J. A.[1], Cardona-Arias, J. A.[2]**

1 Microbiólogo y Bioanalista, Universidad de Antioquia. Grupo de investigación Salud y Sostenibilidad, Escuela de Microbiología, Colombia.
2 Microbiólogo y Bioanalista, MSc Epidemiología. Docente-Investigador Facultad de Medicina, Universidad Cooperativa de Colombia. Escuela de Microbiología, Universidad de Antioquia U de A, Calle 70 No. 52-21, Medellín, Colombia.

**Correspondencia:**
Jaiberth Antonio Cardona-Arias

jaiberthcardona@gmail.com

## Resumen

La información disponible sobre el proceso de construcción, adaptación y validación de escalas en salud es escasa, dispersa y en algunos casos incompleta; por ello el objetivo de este texto es caracterizar los aspectos conceptuales, metodológicos y estadísticos relacionados la construcción y validación de escalas de medición en salud. Se realizó una búsqueda exhaustiva de la literatura referente a la validación de escalas de medición en salud, en google Scholar, Scielo, Science Direct y PubMed/Medline, con aplicación de criterios de inclusión y exclusión. Como resultados centrales se describen las etapas traducción y adaptación de una escala y se presentan las propiedades y estadísticos utilizados para garantizar la reproducibilidad, la validez, la sensibilidad al cambio y la utilidad de las escalas de medición. En la reproducibilidad se describen la fiabilidad, la consistencia interna, el poder discriminante, la fiabilidad test-retest y la fiabilidad inter-observador; mientras que en la validez se incluyen la de apariencia, contenido, criterio (concurrente y predictiva), convergente – divergente y de constructo. Esta revisión representa una síntesis de la información que se encuentra en la literatura sobre la construcción y evaluación de escalas en salud.

**Palabras clave:** Estudios de validación; Escalas; Psicometría; Reproducibilidad de resultados; Validez de las pruebas.

## Abstract

Available information about the process of building, adaptation and validation of scales in health is limited, dispersed and in some cases incomplete; therefore the objective of this text is to characterize the conceptual, methodological and statistical aspects of the construction and validation of measurement scales in health. An exhaustive search of the literature concerning the validation of measurement scales in health, in Google Scholar, Scielo, Science Direct and PubMed/Medline was performed, with application of inclusion and exclusion criteria. The results described the translation and adaptation of a scale and the properties and statistics used to ensure reproducibility, validity, sensitivity to change and usefulness. In the reproducibility are described the properties of reliability, internal consistency, discriminant power, test-retest and inter-observer reliability; while the validity include the appearance, content, criterion (concurrent and predictive), convergent

 

ARCHIVOS DE MEDICINA
ISSN 1698-9465
2015
Vol. 11 No. 3:1
doi: 10.3823/1251

- divergent and construct. This text represents a synthesis of the information found in the literature on the construction and evaluation of health scales.

**Key words:** Validation studies; Scales; Psychometrics; Reproducibility of results; Validity of the tests.

**Fecha de recepción:** Jun 12, 2015, **Fecha de aceptación:** Jul 06, 2015,
**Fecha de publicación:** Jul 14, 2015

# Introducción

Tradicionalmente la medición del estado de salud de los individuos se ha hecho desde la perspectiva biomédica del proceso salud-enfermedad, mediante el uso de marcadores biológicos denominados desenlaces duros u objetivos; sin embargo, teniendo en cuenta la definición de salud de la Organización Mundial de la Salud (OMS) como un estado de completo bienestar físico, mental y social, y no solamente la ausencia de enfermedades, tal abordaje es insuficiente, dado que obsta la multi-dimensionalidad de la salud-enfermedad [1].

El énfasis en la dimensión biofísica se presenta, entre muchas razones, por el hecho de considerar las mediciones no biológicas como indicadores blandos o subjetivos; sin embargo, la perspectiva biomédica resulta limitada, al no incorporar la percepción de salud y calidad de vida propias del paciente, en coherencia con su contexto social, cultural y ambiental [2].

Precisamente, en las áreas de la salud cada vez es más necesario disponer de instrumentos de medida que permitan evaluar atributos subjetivos que integran constructos y dimensiones más complejas, como medio para orientar acciones de atención, promoción o protección de la salud [3]. Tal es el caso de las escalas de medición en salud, diseñadas para evaluar dimensiones físicas, psicológicas o sociales que no pueden observarse ni medirse directamente; cuya importancia radica en que permiten recoger de forma válida y confiable la percepción (subjetiva) del sujeto sobre dichas dimensiones [3-6].

En este sentido, existe un gran número de escalas, tanto genéricas como específicas, que se han desarrollado principalmente en dos campos relacionados: la psicometría y la clinimetría; las cuales son utilizadas ampliamente tanto en la investigación en salud como en la práctica clínica [7]. Por ejemplo, para medir la calidad de vida como fenómeno multidimensional, se han diseñado instrumentos genéricos como el WHOQOL-100, WHOQOL-BREF, MOSSF-36, Duke Health Profile y EUROQOL, empleados en poblaciones sanas y enfermas, con el fin de comparar diferentes entidades clínicas y éstas frente a la población general. Por otro lado, los instrumentos específicos, caracterizados por su sensibilidad clínica, son utilizados para evaluar síntomas, funciones o enfermedades específicas, como son las escalas para personas con Parkinson (PDQ-39), VIH/Sida (MOSHIV), múltiples tipos de cáncer (PROSQOLI), entre otras enfermedades [7-9].

Adicional a lo anterior, se dispone de diversas herramientas psicométricas e instrumentos clinimétricos que permiten evaluar indicadores no biológicos o directamente observables como las habilidades intelectuales, los rasgos de personalidad, la inteligencia, la depresión [3-6], condiciones clínicas como el dolor y la función física en reumatología [10], entre otras.

A pesar de ello, y la elevada cantidad de artículos científicos que han utilizado dichos instrumentos en el ámbito mundial, lo referente al proceso de construcción y validación de escalas de medición en salud sigue presentando limitaciones relacionadas con la falta de claridad en algunas comunidades académicas sobre los criterios que deben evaluarse, la ausencia de consenso sobre el método de construcción y validación de las escalas, y la diversidad de opciones metodológicas con que se llevan a cabo estos procesos [2,3,5,11]. En este orden de ideas, se considera que la ausencia de equivalencia entre las diferentes escalas, derivada de un proceso de validación deficiente, reduce las posibilidad de hacer comparaciones entre poblaciones de diferentes países, culturas e idiomas, impide el intercambio de información en la comunidad científica e induce al diseño de políticas públicas y de salud inadecuadas [11,12].

En este contexto se han desarrollado varias revisiones de tema relacionadas con la validación de escalas en salud como las de Sánchez y Lamprea [2-5], las cuales presentan como limitación el centrarse en asuntos conceptuales, marginar algunas consideraciones metodológicas u operativas y exponer conceptualización poco clara relacionada con algunos criterios de confiabilidad y validez como el poder discriminante y la validez de contenido y constructo. Teniendo en cuenta que la información de la que se dispone actualmente sobre el proceso de validación de escalas es escasa y dispersa, y en muchos casos confusa e incompleta, se realizó una revisión de la literatura con el objetivo de caracterizar los aspectos conceptuales, metodológicos y estadísticos relacionados la construcción y validación de escalas de medición en salud.

# Material y Métodos

## Tipo de studio

Revisión de la literatura.

## Protocolo de investigación

Se realizó una búsqueda exhaustiva de la literatura para localizar la mayor cantidad de información disponible sobre el proceso de validación de escalas de medición en salud. La búsqueda de los artículos se llevó a cabo principalmente en google scholar, con el fin abarcar diversidad de publicaciones referentes al tema e incluso de literatura gris, lo que permitió al mismo tiempo recuperar otras citas sobre artículos relacionados. Así mismo,

2015
Vol. 11 No. 3:1
doi: 10.3823/1251

esta búsqueda fue realizada en las bases de datos Scielo, Science Direct y PubMed/Medline.

Se tomaron como criterios de búsqueda: "validación", "escala", "cuestionario", "instrumento" "estudios de validación", "salud", "medición en salud", "psicometría", "confiabilidad" y "validez".

Los criterios de inclusión para la sección de los artículos fue que éstos estuvieran publicados en español o inglés, que incluyeran las palabras clave o sus combinaciones en el título o resumen, que en lo posible su período de publicación no fuera mayor a 10 años, que trataran los aspectos metodológicos y/o estadísticos del proceso de validación de escalas y que se relacionaran con la medición en salud. Asimismo, se excluyeron los artículos sobre validación de una escala específica y los que no hacían un abordaje teórico, descriptivo o conceptual del proceso de validación.

### Análisis de la información

A partir de los artículos seleccionados se estructuraron los resultados del manuscrito, especificando los principales aspectos conceptuales y metodológicos que deben incluirse en la construcción y validación de escalas de medición en salud, teniendo en cuenta los principales aspectos con los que se pudo establecer un mayor grado de acuerdo entre los autores y sus publicaciones.

## Resultados

Al revisar la literatura sobre la medición en las áreas de la salud, es indiscutible el auge, la demanda y sobretodo la utilidad que tiene la utilización de las escalas para evaluar variables o temas de interés en este campo y cuya característica particular es que no pueden medirse directamente [3-6]. Asimismo, es evidente que estos instrumentos son en su mayoría, traducciones y adaptaciones de versiones originalmente construidas en otro país e idioma, principalmente del inglés [13]. Por lo anterior, y dado que en la literatura existe un amplio consenso sobre cómo abordar esta primera etapa construcción, traducción y adaptación cultural de las herramientas de medida en salud [14,15], la descripción de este proceso no se incluye en el presente artículo, sólo se esquematiza en la **Figura 1**.

En este contexto, suele creerse que la traducción y adaptación de un instrumento ampliamente reconocido y utilizado en un determinado campo, país e idioma garantiza la conservación de sus propiedades psicométricas; sin embargo, esto generalmente no se cumple, por lo que es imperativa su adaptación sociocultural y más importante aún; su validación [12,16].

Vale precisar que el juicio o criterio de expertos es determinante en la etapa inicial de la construcción y validación de escalas en salud, para lo cual existen diversas metodologías y etapas como las propuestas por Escobar y Cuervo en las que se destacan la delimitación del objetivos, los criterios de selección de los expertos, la delimitación de las dimensiones e indicadores a medir, el diseño de la plantilla de evaluación y el cálculo de la concordancia entre jueces [17]. En esta etapa de la validación uno de los métodos más usuales es el Delphi frente al cual existen múltiples publicaciones entre las cuales se destaca la revisión sistemática de García y Suárez en la cual se aluden sus consideraciones conceptuales, se describen su historia, características, ventajas y usos, y se sistematiza su ejecución en el campo de la salud a partir de la delimitación de tres fases: i) preparación, la cual incluye la selección de expertos, preparación del instrumento y decisión de la vía de consulta, ii) consulta, esta incluye la ronda de consultas, el procesamiento estadístico y la realimentación, y iii) consenso y reporte de resultados [18].

Adicional a lo anterior, para considerar válida una escala de medición en salud, ésta debe cumplir con una serie de características como la sencillez, la utilidad (viabilidad), y la aceptación por parte de los pacientes e investigadores, al mismo tiempo que debe satisfacer otros requerimientos íntimamente relacionadas con las dos grandes propiedades psicométricas

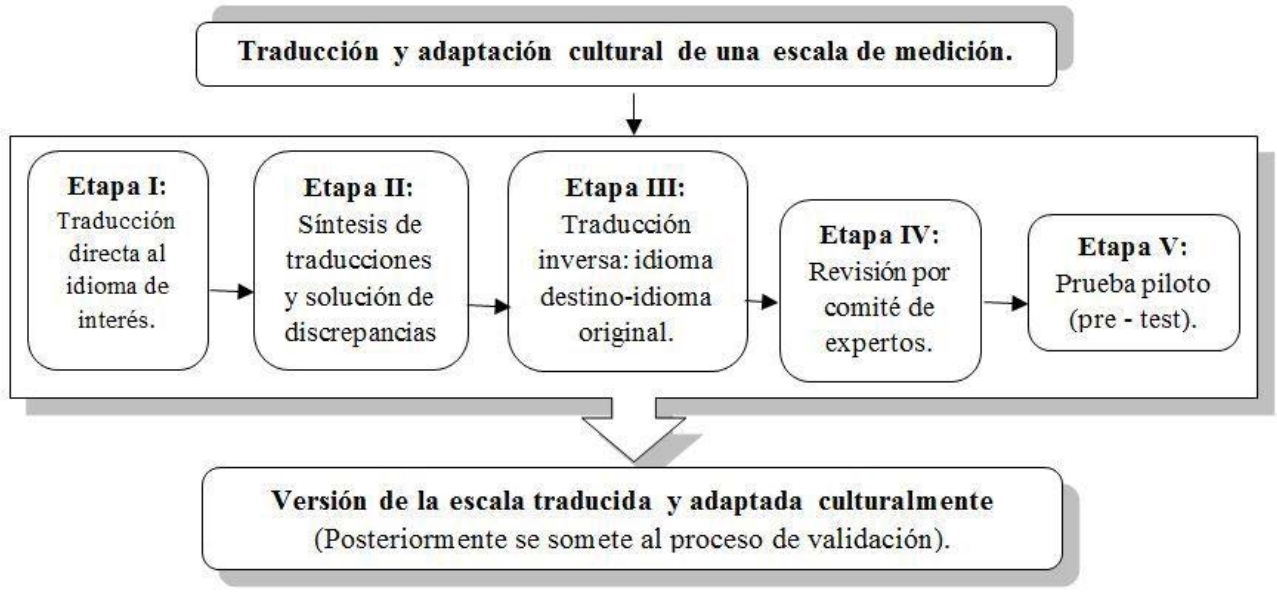


**Figura 1** Síntesis de la metodología sugerida para la traducción y adaptación cultural de una escala. Adaptado de Beaton [14].

determinantes en todo instrumento: la fiabilidad y la validez [19-22].

En este sentido y producto de esta revisión, se sintetiza y propone la siguiente secuencia metodológica y se sugieren algunas herramientas estadísticas para llevar a cabo de manera óptima la validación de una escala de medición en salud, según las propiedades que se resumen en la **Tabla 1.**

## Reproducibilidad

### Fiabilidad

Es el grado en que un instrumento es capaz de medir sin error [23]. Mide la proporción de variación en las mediciones que se debe a la variedad de valores que toma una variable y que no es producto del error sistemático (sesgo) o aleatorio (azar). Es decir, esta propiedad determina la proporción de la varianza total atribuible a diferencias verdaderas que existen entre los sujetos [23,24].

El coeficiente alfa de Cronbach es el recurso estadístico más utilizado para evaluar la fiabilidad de un instrumento [25,26]. Su valor está comprendido entre 0 y 1 y depende tanto del número de ítems que componen la escala como de la correlación media entre ellos [19,27,28]. Adicionalmente, cuando el instrumento está compuesto por un grupo de dominios (sub-escalas), debe calcularse el coeficiente alfa de Cronbach para los ítems de cada dominio respecto del valor del puntaje del mismo (correlación ítem-dominio) [19].

El valor mínimo aceptado para este coeficiente es de 0,70; valores inferiores indican que la fiabilidad de la escala utilizada es baja. Por otro lado, se espera un valor máximo de 0,90; valores mayores indican que hay redundancia o duplicación, lo que significa que varios ítems están midiendo exactamente el mismo elemento de un dominio o constructo; por lo tanto, dichos ítems deben eliminarse. Usualmente, se prefieren valores de alfa entre 0,80 y 0,90 [29].

### Consistencia interna

Es el grado de correlación y coherencia que existe entre los ítems de un instrumento o entre los ítems que conforman una dimensión en las escalas multi-dimensionales [23]. A través de esta propiedad, se evalúa si los ítems que miden una misma dimensión presentan homogeneidad entre ellos, lo que indica que los puntos de cada dominio miden el concepto que pretenden medir y no otro [19,22,30-32]. No obstante, se debe tener en cuenta que las escalas están diseñadas para medir separadamente los diferentes dominios que componen un determinado constructo, por lo cual se debe evaluar la consistencia interna de cada uno de ellos. Una escala cuya consistencia interna es elevada, es decir, aquella en la que sus ítems miden un solo constructo que es homogéneo, garantiza una relación lineal entre la suma de los puntajes de sus ítems con el constructo medido [24,32].

Estadísticamente la consistencia interna se puede evaluar a partir del rango de los coeficientes de correlación de Pearson de cada pregunta con el dominio al cual pertenecen y establecer posteriormente el porcentaje de éxito para cada dominio a partir de la siguiente fórmula:

$$\%\,\text{Éxito Consistencia interna} = \frac{\text{Número de correlaciones ítem – dominio a la cual pertenece} > 0,4}{\text{Número total de correlaciones ítem – dominio a la cual pertenece}} \times 100$$

Adicional a ello, los coeficientes de consistencia interna también pueden desarrollarse por medio del método de división por mitades de Spearman, las de fórmulas de Kuder-Richardson y el α de Cronbach [33]. Una revisión sistemática de la literatura concluye que el α de Cronbach es el coeficiente más utilizado para la evaluación de la consistencia interna de escalas de medición en salud [34]. En este sentido se debe aclarar que en este texto se desagregan dos propiedades relacionadas con la reproducibilidad, estas son fiabilidad y consistencia interna, la primera medida con el α de Cronbach y la segunda con correlaciones.

### Poder discriminante

Esta característica aplica para las escalas multi-dimensionales y determina el grado de correlación que existe entre los ítems de una dimensión y el puntaje de las otras dimensiones a los cuales no pertenecen, es decir, una óptima validez discriminante indica que los ítems de cada dominio no están midiendo lo que los incluidos en las demás dimensiones pretenden medir [32,35].

Para evaluar el poder o la validez discriminante se determina el rango de los coeficientes de correlación de Pearson entre las preguntas y los dominios a los cuales no pertenecen para luego definir el porcentaje de éxito para cada dominio mediante la fórmula:

$$\%\,\text{Éxito Validez discriminante} = \frac{\begin{array}{c}\text{Número de correlaciones ítem – dominio al cual no pertenece} \\ \text{menores que las correlaciones punto – dominio al cual pertenece}\end{array}}{\text{Número total de correlaciones del punto – dominio al cual no pertenece}} \times 100$$

### Fiabilidad intra-observador o fiabilidad test-re-test

Se refiere a la repetibilidad del instrumento, es decir, si cuando es aplicado por los mismos evaluadores, con el mismo método, a la misma población y en dos momentos diferentes se obtienen puntajes similares [19,23]. Para evaluar esta propiedad se puede usar el coeficiente de correlación de Pearson, Spearman o intraclase [11]. El coeficiente de Pearson, se utiliza para medir la correlación entre variables cuantitativas al igual que el CCI, sin embargo, si las variables son cualitativas ordinales, está más indicada la correlación de Sperman-Brown [36-38].

Además, es importante considerar que el tiempo transcurrido entre la primera aplicación de la escala (test) y la segunda (retest) varía según lo que se esté midiendo. Esto es, no debe ser largo, para evitar variaciones en el fenómeno de interés medido y tampoco debe ser muy breve, ya que puede presentarse un "efecto de aprendizaje", es decir, recordar las respuestas dadas en la primera aplicación. En los dos casos, el valor de la repetibilidad se ve alterado. Otra dificultad que se puede presentar es que algunos sujetos no admitan una segunda aplicación del instrumento [12,37,38].

### Fiabilidad inter-observador

Se refiere al grado de acuerdo que hay entre evaluadores diferentes que valoran a los mismos sujetos, con el mismo instrumento y en la misma ocasión. Esta propiedad no es

**Tabla 1** Resumen de una evaluación psicométrica de escalas en salud.

| Criterio | Propiedad | Definición | Estadístico | Resultado satisfactorio |
|---|---|---|---|---|
| **Reproducibilidad** | Fiabilidad | Variación u homogeneidad en las mediciones | Coeficiente alfa de Cronbach | ≥ 0,7 |
| | Consistencia interna | Correlación entre los ítems de una dimensión (aplica para escalas multidimensionales e índices) | Correlación de Pearson, Spearman o Kuder-Richardson | ≥ 0,4 (en caso de ser ≥0,9 indicaría mediciones son iguales) |
| | Poder discriminante | Correlación entre los ítems de una escala y las dimensiones a las cuales no pertenecen (sólo en escalas multidimensionales) | Correlación de Pearson o Spearman | Menor a la correlación del ítems con su dimensión (<0,3) |
| | Fiabilidad intra-observador o test-retest | Repetibilidad del instrumento | Correlación de Pearson, Spearman o intraclase | ≥0,80 ó 0,85 |
| | Fiabilidad inter-observador | Concordancia en evaluadores diferentes con los mismos sujetos, igual instrumento y ocasión | Correlación de Pearson, Spearman o intraclase | ≥0,80 ó 0,85 |
| **Validez** | De apariencia (lógica) | Grado en que los ítems mide de forma lógica un constructo dado | Ninguno. Aplicabilidad y aceptabilidad | No aplica |
| | De contenido | Los ítems del instrumento representan adecuadamente el constructo que pretende medir | Análisis factorial exploratorio | Coeficientes λ o cargas factoriales ≥0,3 |
| | De criterio (concurrente y/o predictiva) | Grado de similitud en los puntajes de la escala comparados con un estándar o patrón de referencia (criterio) | Coeficientes de correlación de Pearson o de Spearman | ≥0,80 |
| | Convergente / divergente | Correlaciona los puntajes obtenidos con escalas diferentes | Correlación de Pearson o de Spearman | Entre 0,4 y 0,70 |
| | De constructo | Grado en que el instrumento refleja adecuadamente la teoría subyacente del fenómeno o constructo que se quiere medir | Análisis factorial confirmatorio. O pruebas de hipótesis para comparar grupos teóricamente diferentes | Coeficientes λ ≥0,3, estadísticos de bondad de ajuste ≥0,05. En pruebas de hipótesis Vp<0,05 |
| **Sensibilidad** | Capacidad de un instrumento para detectar cambios a través del tiempo | | Pruebas de hipótesis | Vp <0,05 |
| **Utilidad** | La escala es de fácil aplicación, poca compleja y bajo costo | | Ninguno | No aplica |

evaluable en instrumentos donde el mismo individuo es el que proporciona las respuestas (test autocompletados), sin que exista interferencia de los evaluadores en los resultados del mismo (ej. entrevista). Las limitaciones con esta medición se deben principalmente a la existencia de acuerdo entre los evaluadores por azar y a la presencia de error sistemático (sesgo de información) en alguno de ellos [12,19]. Si se precisa su evaluación, los métodos estadísticos usados son los mismos que para la fiabilidad test-retest.

## Validez

La validez es la capacidad que tiene el instrumento para medir el constructo que pretende medir y para lo cual fue diseñado [23]. Se reconocen cinco que componen la validez de un instrumento: validez de apariencia, de contenido, de criterio, convergente-divergente y de constructo [23,24]. La validez podrá evaluarse para todas o algunas de estas dimensiones dependiendo del tipo de escala objeto de la validación.

## De apariencia (lógica)

Hace referencia al grado en que los ítems (preguntas) de una escala, mide de forma aparente o lógica el constructo que se pretende medir [23]. Para evaluar esta propiedad deben conformarse dos grupos, uno de expertos y otro de sujetos que serán medidos con el instrumento. Ambos analizan la escala y deciden si las preguntas realmente parecen medir lo que se quiere. Cabe aclarar que la validez de apariencia no es un concepto estadístico, sino que depende del juicio que hagan los expertos sobre la conveniencia de los ítems para evaluar el constructo de interés [5]. Además, la relevancia de esta forma de

validez reside en la aplicabilidad y sobre todo en la aceptabilidad desde el punto de vista de quien responde y es evaluado con la escala [39].

## Validez de contenido

Esta propiedad busca evaluar si los diferentes ítems incluidos en el instrumento representan adecuadamente los dominios del constructo que se pretende medir [4,19,23]. La validez de contenido es un proceso en el que se determina la estructura de la escala garantizando que ésta, por medio de sus ítems, abarque todos los dominios de la entidad que se quiere medir, es decir, confirmar que el fenómeno estudiado esté representado adecuada y totalmente por sus ítems y dominios sin dejar ningún aspecto fuera de la medición lo que significa que abarca el espectro real de la entidad, de tal modo que las inferencias surgidas a partir del puntaje de la escala sean válidas dentro de un amplio rango de circunstancias [4,19].

El procedimiento para evaluar la validez de contenido supone aplicar métodos estadísticos como el análisis factorial exploratorio [40,41], éste se usa para obtener evidencias de las dimensiones subyacentes (componentes) que están presentes en el instrumento y que deberían corresponder, en teoría, al constructo que se quiere medir. Con esto se busca explicar las correlaciones existentes entre los ítems del instrumento a partir de un conjunto más pequeño de componentes llamados dominios o "factores"; en este análisis es determinante evaluar el ajuste del modelo factorial y la adecuación de la muestra y los ítems evaluados, para lo cual se utilizan el test de esfericidad de Barlet y el de Kaiser-Meyer-Olkin (KMO), este último se toma como satisfactorio para valores mayores a 0,7; adicional a las rotaciones, principalmente la ortogonal varimax [6,40,41]. A nivel global, las cargas o saturaciones factoriales de los ítems (correlación entre cada ítem y cada factor) se consideran óptimas si son iguales o mayores a 0,3 [6].

## Validez de criterio

Establece el grado en que los puntajes obtenidos a partir de una escala son válidos, al compararlo con un estándar o patrón de referencia (criterio) [3,19,23]. En este caso, el nuevo instrumento que se está evaluando debe compararse con una escala existente que sea ampliamente aceptada y haya demostrado ser el mejor instrumento disponible para la medición del fenómeno de interés. De este modo, se comparan los puntajes obtenidos con cada una de las escalas con el fin de evaluar si existe una adecuada correlación entre ambas [4,19].

Siempre que exista un estándar o se disponga de una escala alternativa que haga sus veces y que además sea independiente, fiable, válida y por supuesto, que mida la misma condición de interés, se deben seguir los siguientes pasos para evaluar esta propiedad: seleccionar el estándar o su equivalente más adecuado, elegir una muestra representativa de la población objeto de estudio, aplicar la escala en evaluación y obtener un puntaje para cada individuo, evaluar a cada sujeto con el estándar y comparar los resultados obtenidos con ambos instrumentos [12,19].

Dependiendo del momento en que se realice la comparación de resultados, pueden evaluarse las características de esta propiedad: la validez concurrente y la validez predictiva [4,42]. La validez concurrente busca establecer el grado de correlación existente entre los resultados obtenidos por la escala en evaluación y la considerada "criterio" o estándar, cuando ambas son aplicadas simultáneamente [4]. Esta comparación se efectúa estadísticamente mediante coeficientes de correlación de Pearson [3,19] o de Spearman, dependiendo de las características de distribución de los datos [42].

La validez predictiva evalúa el grado en que la nueva escala de medición es capaz de predecir el puntaje obtenido por el estándar de oro cuando éste no se aplica al mismo tiempo sino en algún punto en el futuro [4]. Estadísticamente, esta comparación se realiza de igual forma que en la validez concurrente.

La finalidad de la validez de criterio es que exista una adecuada correlación entre ambos instrumentos. Vale precisar que si la finalidad de la validación de una nueva escala está fundamentada en que ésta presenta mayor utilidad, ya sea por la simplicidad en su aplicación, calificación, comodidad para el individuo, economía, disminución del error de medida y pertinencia; la validación de criterio requiere la obtención de correlaciones iguales o mayores a 0,8 las cuales indican que las dos escalas son psicométricamente iguales. Si la finalidad es mostrar que la nueva escala es más válida y mejor que el instrumento de referencia, lo ideal sería obtener correlaciones entre 0,3 y 0,7; las cuales indican que los dos instrumentos son diferentes aunque miden el mismo atributo [3,39,42].

## Validez de constructo

Garantiza que los puntajes que resultan de las respuestas del instrumento puedan ser consideradas y utilizadas como una medición válida del fenómeno estudiado [19,23]. Así, esta propiedad evalúa el grado en que el instrumento refleja adecuadamente la teoría subyacente del fenómeno o constructo que se quiere medir y en consecuencia, la medida coincide con la de otros instrumentos que evalúan la misma condición [6,19,23,40,41]. La evaluación de estos atributos o constructos demanda la definición previa del contenido del instrumento que se está validando y la elaboración de un marco teórico-conceptual que permita la interpretación los resultados obtenidos. De este modo, la validez de constructo permite establecer cómo una medición de la entidad se relaciona de manera consistente con las hipótesis que se plantean para explicar el constructo teórico que define el fenómeno de interés [6,19,40,41].

Estadísticamente, la evaluación de esta propiedad se hace mediante análisis factorial, precisando que inicialmente se usa el análisis factorial exploratorio para revelar la estructura interna de ítems y factores (dominios) de la escala y posteriormente, el análisis factorial confirmatorio para dar validez a tal estructura factorial soportada en un marco teórico de referencia. Aunque también es posible utilizar pruebas de hipótesis para comparar grupos teóricamente diferentes y con ello, evidenciar que la escalas los discrimina adecuadamente y en consecuencia, el constructo es válido [6,11,40,41,43].

El análisis factorial confirmatorio (AFC) es la herramienta estadística más apropiada para evaluar empíricamente la configuración teórica (constructo) subyacente de un instrumento, en términos de las características o rasgos latentes que representa, incluidos sus ítems y factores dentro de una posible estructura jerárquica [6,44]. Es decir, el AFC otorga evidencia suficiente de la validez del constructo permitiendo así, contrastar y evaluar hipótesis correctamente formuladas y validar las deducciones teóricas inferidas del mismo a la luz de los puntajes que se obtienen con la escala [40,41,43,44].

### Validez convergente/divergente

Esta propiedad correlaciona los puntajes obtenidos a través escalas diferentes. Si se comparan instrumentos que cuantifican el mismo constructo y los resultados entre ambas medidas presentan correlaciones significativas, se dice que "convergen", lo cual comprueba que las escalas son conceptualmente congruentes o similares. Si por el contrario, se comparan los puntajes de escalas que miden constructos diferentes y se obtienen correlaciones bajas o negativas, significa que las escalas "divergen", indicando asociación no significativa entre las variables, lo que confirma que miden constructos distintos, de no ser así, significaría que la escala que se está validando no es lo suficientemente específica para medir el constructo de interés en una población dada [11,22,30,45,46].

Esta propiedad generalmente se evalúa con coeficientes de correlación de Pearson, esperando obtener en circunstancias ideales, un desempeño psicométrico aceptable que implica una correlación mayor de 0,60 (convergente) y una correlación menor de 0,20 (divergente) [47,48].

### Sensibilidad

La sensibilidad es la capacidad de un instrumento para detectar cambios a través del tiempo en la realidad que mide, tanto entre los individuos como en la respuesta de un mismo individuo sobre dicho constructo. Esta propiedad es común en escalas diagnósticas, ensayos clínicos o mediciones prospectivas, en los que la sensibilidad al cambio y la especificidad permiten evaluar la respuesta a un tratamiento o intervención; sin embargo, es poco frecuente en estudios con variables como el bienestar, la satisfacción, las percepciones y las actitudes [5,19,23,24,49].

### Utilidad

Un instrumento no es útil si su aplicación resulta difícil, compleja o costosa. Este parámetro hace referencia a aspectos como el tiempo necesario para la aplicación del instrumento, la sencillez en el formato, la claridad de las preguntas, si se requiere o no de entrenamiento al personal que lo aplica. Además identifica si su registro, codificación, interpretación y evaluación es simple. Esta característica se evalúa mediante la realización de una prueba piloto, con grupo pequeño de participantes, de modo tal que puedan realizarse modificaciones oportunas en términos de su viabilidad [5,12,19].

## Discusión

Esta revisión pone de manifiesto el creciente interés por desarrollo y uso de escalas de medición en salud, con un amplio número de artículos referidos a la validación de instrumentos de reconocida utilidad y pertinencia en aéreas como la educación, la psicología, las ciencias sociales y de la salud [3-6]. Sin embargo, no sucede lo mismo con las investigaciones acerca de la sistematización y descripción metodológica del proceso que implica la validación completa, rigurosa y exhaustiva de estas herramientas de medida.

En este sentido, se pudo observar que si bien muchos autores se han interesado por investigar y describir estos aspectos metodológicos desde un enfoque teórico fundamentado principalmente desde la perspectiva psicométrica y clinimétrica [3-8], es clara la falta de consenso en aspectos fundamentales como la terminología, definiciones, opciones estadísticas, criterios psicométricos mínimos a evaluar y en general, los procesos de construcción, adaptación y validación de los instrumentos [23].

En relación con la fiabilidad y la validez vale precisar que, si bien muchos estudios las incluyen, en especial la medición del alfa de Cronbach y correlaciones de Pearson [20-22], los estudios que sólo velan por estas propiedades deberían ser más exhaustivos en los tipos de validación que subsumen; además de tener presente que el cumplimiento de éstas no garantiza la equivalencia entre las diferentes versiones que han sido desarrolladas para su aplicación en países con idiomas y culturas diferentes. Es determinante la aplicación de las propiedades descritas para realizar comparaciones entre estas poblaciones, obtener resultados validos derivados y con ello, la posibilidad de generar perfiles poblacionales, políticas públicas adecuadas, tratamientos eficientes e intervenciones en salud coherentes con la realidad de dichos grupos [12-16].

Cabe aclarar que las propiedades presentadas no aplican de forma similar para todos los instrumentos y cada investigador, según el tipo de escala que esté evaluando (unidimensional o multidimensional, con diversas variables latentes o sin ellas, adaptadas culturalmente o no, adecuadamente traducidas o no) determinará las propiedades que apliquen para su estudio. En este orden de ideas vale citar algunos estudios previos que han aplicado algunas propiedades psicométricas descritas:

1. El grupo de Mayta P. diseñó y evaluó una escala para medir percepciones sobre el trabajo en el primer nivel de atención de estudiantes de medicina, para lo cual realizaron un análisis factorial exploratorio de componentes principales y se estimó la varianza explicada [50].

2. Quintero C. y su grupo realizó la validación del cuestionario KIDSCREEN-27 de calidad de vida de niños y adolescentes en Medellín, para lo cual evaluaron las propiedades de consistencia interna, fiabilidad inter e intra-observador, validez de contenido y de constructo, y sensibilidad al cambio [51].

3. El grupo Cardona J, evaluó la fiabilidad, consistencia interna, validez discriminante y validez convergente/divergente de tres escalas de calidad de vida para personas con fibromialgia, FIQ (Fibromyalgia Impact Questionnaire), MOSSF-36 (Medical Outcome Study Short Form) y WHOQOL-BREF (World Health Organization Quality of Life) [52], para personas con VIH se evaluaron

propiedades psicométricas similares para las escalas MOSSF-36, WHOQOL-BREF y WHOQOL-HIV-BREF [53] y en población sana se evaluó validez discriminante, convergente/divergente, fiabilidad y consistencia interna del WHOQOL-BREF y el MOSSF-36 [54].

4. El grupo de Hernández evaluó la fiabilidad, la consistencia interna y validez discriminante y predictiva de una encuesta de frecuencia de consumo de alimentos ricos en hierro en Medellín-Colombia [55].

## Conclusiones

Esta revisión representa una síntesis y organización de la información que se encuentra ampliamente dispersa en la literatura en relación con la evaluación psicométrica de escalas en salud, al mismo tiempo que hace una aproximación a los recursos estadísticos que más se utilizan para lograr este propósito. Por lo anterior, este trabajo resulta de gran utilidad para investigadores de las áreas de salud y afines, interesados en construir, adaptar y evaluar los instrumentos de su interés, de modo que los resultados de sus investigaciones sean reproducibles, válidos, sensibles y útiles para la población estudiada y la comunidad científica.

## Financiación

Recursos en especie Universidad de Antioquia, Universidad Cooperativa de Colombia. Estrategia de sostenibilidad 2014.

## Conflictos de Interés

Los autores declaran no tener conflictos de interés con la publicación de este artículo.

## Bibliografia

1 World Health Organization (WHO). Official records of the World Health Organization. International Health Conference. [Internet] 1946. [Consulta 10 May 2015]. Disponible en: http://whqlibdoc.who.int/hist/official_records/2e.pdf.

2 Grupo de la Organización Mundial de la Salud sobre la calidad de vida. Que calidad de vida? Foro Mundial de la Salud. Rev Inter Desar Sanit. 1996; 17: 385-7.

3 Sánchez, R., Gómez, C. Conceptos básicos sobre la validación de escalas. Rev. Col. Psiquiatría. 1998; 27: 121-30.

4 Lamprea, J, Gómez, C. Validez en la evaluación de escalas. Rev. Colomb. Psiquiat. 2007; 36: 340-8.

5 Sánchez, R., Echeverry, J. [Validating scales used for measuring factors in medicine]. Rev Salud Publica (Bogota) 2004; 6: 302-318.

6 Montero E. Referentes conceptuales y metodológicos sobre la noción moderna de validez de instrumentos de medición: implicaciones para el caso de personas con necesidades educativas especiales. Actualidades en Psicología. 2013; 27: 113-28.

7 Lara, M., Ortega, H. Psicometría o clinimetría? Medición en la práctica psiquiátrica. Salud mental. 1995; 18: 33-40.

8 Velarde-Jurado, E., Avila-Figueroa, C. [Methodological considerations for evaluating quality of life]. Salud Publica Mex 2002; 44: 448-463.

9 Guyatt, GH.., Feeny, DH., Patrick, DL. Measuring health-related quality of life. Ann Intern Med 1993; 118: 622-629.

10 Schneeberger, E., Marengo, M., Papasidero, S., Chaparro, R., Citera, G. Clinimetría en Artritis Reumatoidea. Revista Argentina de Reumatología. 2008; 19: 3-26.

11 Carvajal, A., Centeno, C., Watson, R., Martínez, M., Rubiales, AS. [How is an instrument for measuring health to be validated?]. An Sist Sanit Navar 2011; 34: 63-72.

12 Ramada-Rodilla, JM., Serra-Pujadas, C., Delclós-Clanchet, GL. [Cross-cultural adaptation and health questionnaires validation: revision and methodological recommendations]. Salud Publica Mex 2013; 55: 57-66.

13 Elosua, P. Tests publicados en España: usos, costumbres y asignaturas pendientes. Papeles del psicólogo. 2012; 33: 12-21.

14 Beaton, DE., Bombardier, C., Guillemin, F., Ferraz, MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1976) 2000; 25: 3186-3191.

15 Guillemin, F. Cross-cultural adaptation and validation of health status measures. Scand J Rheumatol 1995; 24: 61-63.

16 Muñiz, J., Elosua, P., Hambleton, RK; International Test Commission [International Test Commission Guidelines for test translation and adaptation: second edition]. Psicothema 2013; 25: 151-157.

17 Escobar, J., Cuervo, A. Validez de contenido y juicio de expertos: Una aproximación a su utilización. Avances de Medición. 2008; 6: 27-36.

18 García, M., Suárez, M. El método Delphi para la consulta de expertos en la investigación científica. Rev Cub Salud Pública. 2013; 39: 253-67.

19 García de Yébenes Prous, MA., Rodríguez Salvanés, F., Carmona Ortells, L. [Validation of questionnaires]. Reumatol Clin 2009; 5: 171-177.

20 Prieto, G., Delgado, A. Fiabilidad y validez. Papeles del Psicólogo. 2010; 31: 67-74.

21 Quintana, A., Montgomery, W. Psicología: Tópicos de actualidad. Lima-Perú: UNMSM. 2006; 86-88.

22 Argibay, J. Técnicas psicométricas. Cuestiones de validez y confiabilidad. Subjetividad y procesos cognitivos. 2006; 15-33.

23 Mokkink, LB., Terwee, CB., Patrick, DL., Alonso, J., Stratford, PW., et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010; 19: 539-549.

24 Argimon, J., Jiménez, V. Métodos de investigación clínica y epidemiológica. 3a. ed. Madrid: Elsevier España. 2004.

25 Soler, S. Coeficientes de confiabilidad de instrumentos escritos en el marco de la teoría clásica de los tests. Educ Med Super. 2008; 22: 1-14.

26 Cronbach, L. Coefficient alpha and internal structure of test. Psychometrika.1951; 16: 297-333.

27 Bland, JM., Altman, DG. Cronbach's alpha. BMJ 1997; 314: 572.

28 Soler S, Soler L. Usos del coeficiente alfa de Cronbach en el análisis de instrumentos escritos. Rev. Med. Electrón. 2012; 34: 1-6.

29 Streiner, DL Starting at the beginning: an introduction to coefficient alpha and internal consistency. J Pers Assess 2003; 80: 99-103.

30 Carretero, H., Pérez, C. Normas para el desarrollo y revisión de estudios instrumentales. International Journal of Clinical and Health Psychology. 2005; 5: 521-51.

31 Oviedo, H., Campo, A. Aproximación al uso del coeficiente alfa de Cronbach. Rev. Colomb. Psiquiatr. 2005; 34: 572-80.

32 Campo-Arias, A., Oviedo, HC. [Psychometric properties of a scale: internal consistency]. Rev Salud Publica (Bogota) 2008; 10: 831-839.

33 Aiken, L. Test Psicológicos y evaluación. Confiabilidad y validez. Undécima edición. México: Pearson Educación. 2003; 85-107.

34 Cascaes da Silva, F., Gonçalves, E., Valdivia Arancibia, BA., Bento, GG., Silva Castro, TL., et al. [Estimators of internal consistency in health research: the use of the alpha coefficient]. Rev Peru Med Exp Salud Publica 2015; 32: 129-138.

35 Martínez, J., Martínez, L. La validez discriminante como criterio de evaluación de escalas: teoría o estadística? Universitas Psychologica. 2009; 8: 27-36.

36 Fortin, M., Nadeau, M. La medida de investigación. Fortin MF (Ed). El proceso de investigación de la concepción a la realización. México: McGraw-Hill Interamericana. 1999.

37 Prieto, L., Lamarca, R., Casado, A. [Assessment of the reliability of clinical findings: the intraclass correlation coefficient]. Med Clin (Barc) 1998; 110: 142-145.

38 Müller, R., Büttner, PA. critical discussion of intraclass correlation coefficients. Stat Med 1994; 13: 2465-2476.

39 Streiner, DL. A checklist for evaluating the usefulness of rating scales. Can J Psychiatry 1993; 38: 140-148.

40 Pérez, A., Chacón, M., Moreno, R. Validez de constructo: el uso de análisis factorial exploratorio-confirmatorio para obtener evidencias de validez. Psicothema. 2000; 12: 442-6.

41 Batista-Foguet, JM., Coenders, G., Alonso, J. [Confirmatory factor analysis. Its role on the validation of health related questionnaires]. Med Clin (Barc) 2004; 1: 21-27.

42 Streiner, D., Norman, G. Health Measurement Scales. A Practical Guide to their Development and Use. 2nd ed. Oxford: Oxford University Press. 1995.

43 Zamora, S., Monroy, L., Chávez, C. Análisis factorial: una técnica para evaluar la dimensionalidad de las pruebas. Cuaderno técnico 6. México D.F.: Centro Nacional de Evaluación para la Educación Superior, A.C. 2010.

44 Brown, T. Confirmatory Factor Analysis for Applied Research. New York: the Guilford Press. 2006.

45 Alarcón Ma, AM., Muñoz, NS. [Some methodological issues about measurements in health]. Rev Med Chil 2008; 136: 125-130.

46 Blacker, D., Endicott, J. Psychometric properties: concepts of reliability and validity. En: Rush A, First M, Blacker D. Handbook of psychiatric measures. 2da Ed. Washington: APA; 2007.

47 DeVon, HA., Block, ME., Moyle-Wright, P., Ernst, DM., Hayden, SJ., et al. A psychometric toolbox for testing validity and reliability. J Nurs Scholarsh 2007; 39: 155-164.

48 Pita, S., Pértegas, J. Relación entre variables cuantitativas. Cad Aten Primaria. 1997; 4: 141-144.

49 Guyatt, G., Walter, S., Norman, G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987; 40: 171-178.

50 Mayta-Tristán, P., Mezones-Holguín, E., Pereyra-Elías, R., Montenegro-Idrogo, JJ., Mejia, CR, Dulanto-Pizzorni, A., et al. Diseño y validación de una escala para medir la percepción sobre el trabajo en el primer nivel de atención en estudiantes de medicina de Latinoamérica. Rev Peru Med Exp Salud Publica. 2013; 30: 190-196.

51 Quintero, C., Lugo, L., García, H., Sánchez, A. Validación del cuestionario KIDSCREEN-27 de calidad de vida relacionada con la salud en niños y adolescentes de Medellín, Colombia. Rev. Colomb. Psiquiatr. 2011; 40: 470-487.

52 Cardona, J., Hernández, A. León, V. Validez, fiabilidad y consistencia interna de tres instrumentos de medición de calidad de vida relacionada con la salud en personas con fibromialgia, Colombia. Rev.Colomb.Reumatol 2014; 21: 57-64.

53 Cardona, J. Calidad de vida relacionada con la salud en personas con VIH/SIDA: Comparación del MOSSF-36, WHOQOL-BREF y WHOQOL-HIV-BREF, Medellín-Colombia Colombia Médica. 2011; 42: 438-47.

54 Cardona J, Ospina L, Eljadue A. Validez discriminante, convergente/divergente, fiabilidad y consistencia interna, del WHOQOL-BREF y el MOSSF-36 en adultos sanos de un municipio colombiano. Rev fac Nac Salud Pública. 2015; 33: 50-57.

55 Hernández, A., Mantilla, C., Cardona, J. Evaluación de la validez y fiabilidad de una encuesta de frecuencia de consumo de alimentos ricos en hierro, Medellín-Colombia 2013. Archivos de Medicina. 2014; 10: 20.