Artículo especial

Criterios útiles para la selección de instrumentos de medición en disciplinas de la salud

Key Criteria for Selecting Measurement Instruments in Health Disciplines

Francisco Javier Fulvio Gómez-Clavelina[1] Geovani López-Ortiz[1]*

Resumen

Antecedentes: se ha identificado que la validez y la confiabilidad, en diversos tipos de estudio, con frecuencia se malinterpretan y se aplican de forma inadecuada. Por tal motivo, es indispensable que los profesionales de la salud cuenten con instrumentos adecuados para la medición de atributos subjetivos, así como de dimensiones complejas. Esto permitirá una adecuada interpretación de la evidencia y una mejora en la atención médica. **Objetivo:** proponer elementos útiles para la selección adecuada de instrumentos de medición en el ámbito de la salud, que permitan evaluar variables de interés. Métodos: revisión de literatura. Mediante algoritmos de búsqueda, en diferentes bases de datos, se seleccionaron revisiones narrativas y sistemáticas publicadas de 2000 a 2023 que incluyeran aspectos relacionados con instrumentos de medición. Resultados: de 351 artículos identificados en un primer análisis, se seleccionaron 62 que cumplieron con los criterios de selección. Los elementos que se consideraron para la elección de instrumentos de medición fueron: adaptación cultural, fiabilidad, confiabilidad, validez de constructo teórico, validez de constructo empírico (criterio y constructo) y capacidad de respuesta. De todos estos elementos se obtuvieron los aspectos más importantes en su definición, clasificación y utilidad. **Conclusiones:** para seleccionar adecuadamente los instrumentos de medición de variables cuyas características no permiten una medición directa, es fundamental la búsqueda y evaluación de las propiedades descritas en este artículo, las cuales proporcionan una manera más objetiva de seleccionarlos, permitiendo que se realicen mediciones más precisas del fenómeno a evaluar, tanto desde la visión cualitativa como cuantitativa.

Palabras clave: validez, confiabilidad, psicometría, medición, investigación biomédica, evaluación, diseño de investigación.

Recibido: 02/08/2024
Aceptado: 22/11/2024

[1]Subdivisión de Medicina Familiar, División de Estudios de Posgrado, Facultad de Medicina, Universidad Nacional Autónoma de México.

*Correspondencia:
Francisco Javier Fulvio Gómez-Clavelina
igc@unam.mx

Sugerencia de citación: Sugerencia de citación: Gómez-Clavelina FJF, López-Ortiz G. Criterios útiles para la selección de instrumentos de medición en disciplinas de la salud. Aten Fam;32(2):127-137.http://dx.doi.org/10.22201/fm.14058871p.2025.2.91029

Summary

Background: Validity and reliability are frequently misinterpreted and improperly applied in various types of studies. Therefore, it is essential for healthcare professionals to use appropriate measurement instruments for assessing subjective attributes and complex dimensions. This ensures accurate interpretation of evidence and improves medical care. **Objective:** To propose useful criteria for selecting appropriate measurement instruments in the health field to evaluate variables of interest. **Methods:** A literature review was conducted. Using search algorithms in various databases, narrative and systematic reviews published between 2000 and 2023 that addressed measurement instruments were selected. **Results:** Of the 351 articles initially identified, 62 met the selection criteria. The key criteria considered for selecting measurement instruments included cultural adaptation, reliability, validity of theoretical constructs, validity of empirical constructs (criterion and construct), and responsiveness. The most relevant aspects of these criteria were defined, classified, and analyzed in terms of their usefulness. **Conclusions:** Selecting measurement instruments for variables that cannot be directly measured requires a thorough search and evaluation of the properties described in this article. These criteria provide a more objective approach to selecting measurement tools, enabling more precise assessments of the phenomena under study from both qualitative and quantitative perspectives.

Key words: Validity; Reliability; Psychometric; Measurement; Biomedical research; Assessment; Research design.

Introducción

En el desempeño cotidiano de sus actividades, los profesionales de la salud deben realizar mediciones de muy diversos elementos vinculados con la salud de los usuarios de los servicios de atención sanitaria.

La perspectiva biomédica del proceso salud-enfermedad suele ser la que predomina y caracteriza la visión cuantitativa de variables como la tensión arterial, la estatura, los niveles de hemoglobina o colesterol en sangre, etc. Sin embargo, el campo de acción de los profesionales de la salud incluye también aspectos vinculados con la salud mental, el contexto social, cultural y ambiental. Dada la naturaleza multidimensional del proceso salud-enfermedad, la visión biomédica cuantitativa es insuficiente.[1]

Las variables de interés como la calidad de vida, el resultado de una intervención, la inteligencia, la depresión o el estrés, suelen ser abstracciones que representan fenómenos o constructos que no pueden observarse directamente.[2] Para lograr la medición de este tipo de variables, se hace necesario utilizar indicadores observables como las respuestas a preguntas de un cuestionario o la declaración de síntomas o percepciones de los sujetos de estudio. La fidelidad de esta medición depende de la relación entre estos indicadores observables y los constructos subyacentes. Si la relación es débil, las inferencias efectuadas serán imprecisas y probablemente incorrectas.[3]

Es indispensable que los profesionales de la salud cuenten con instrumentos diseñados para la medición, tanto de atributos subjetivos como la evaluación de dimensiones complejas, con lo cual se logre una adecuada orientación y toma de decisiones en la atención, promoción y protección de la salud.[4,5] Con este propósito se han desarrollado al menos dos enfoques para evaluar las mediciones y los instrumentos que se utilizan para obtenerlas: la clinimetría y la psicometría.[6]

Se ha identificado que la validez y la confiabilidad con frecuencia se malinterpretan.[7] El número de publicaciones científicas en las que se refiere el uso de algún instrumento de medición (test, cuestionario, índice, escala, etc.) es enorme y los grupos académicos respaldan la fiabilidad y validez de las mediciones con procedimientos diversos basados ya sea en la clinimetría o en la psicometría; no obstante, es necesario resaltar que, estudios de investigación con métodos sólidos a menudo no presentan un amplio espectro de evidencia de validez que respalde los resultados.[8] Esto permite inferir la carencia de consenso acerca de los métodos de construcción y validación de los instrumentos de medición.[1,9]

Dado lo anterior, el objetivo de este estudio es presentar una propuesta de elementos que los investigadores y clínicos en el ámbito de las disciplinas de la salud, deben considerar para seleccionar adecuadamente un instrumento que permita la medición de variables de interés.

Métodos

En esta revisión se consultaron las siguientes bases de datos: PubMed/Medline, Scielo, *Science Direct, Medscape, Research Gate, Excerpta Medica Database (EMBASE), Health Management Information Consortium (HMIC) y Google Scholar.*

La búsqueda incluyó artículos de enero de 2000 a julio de 2023. Se utilizaron los siguientes términos MeSH (*Medical Subject Headings*): fiabilidad, validez, escala, cuestionario, estudios de validación, medición en salud, psicometría, clinimetría, metodología, investigación médica y sus traducciones

en inglés. Criterios de búsqueda: a. Artículos en inglés y español; b. Que ocuparon los términos MeSH o sus combinaciones en el título o resumen; c. Que trataron los aspectos metodológicos y/o estadísticos del proceso de validación de escalas; d. Que se relacionaron con la medición en salud y e. De acceso libre, artículos de revisión y revisiones sistemáticas. Se excluyeron comentarios, editoriales y documentos que no hacían un abordaje teórico, descriptivo o conceptual del proceso de validación (figura 1).

Análisis de la información

Se revisaron los resúmenes en las bases de datos y con fundamento en los criterios de selección se descargaron los artículos. Con los contenidos de los artículos seleccionados, se estructuraron los resultados de la revisión. Debido a la gran diversidad de criterios y definiciones vertidas en las publicaciones, se optó por seleccionar las categorías y subcategorías de la psicometría, la clinimetría, y las iniciativas internacionales identificadas mediante la revisión de los artículos.

Resultados

Se descargaron 351 artículos; la revisión de sus resúmenes permitió organizar los contenidos temáticos e identificar las diferencias en términos, conceptos y metodologías propuestas por los autores. Las fuentes finales de información fueron 62 artículos.

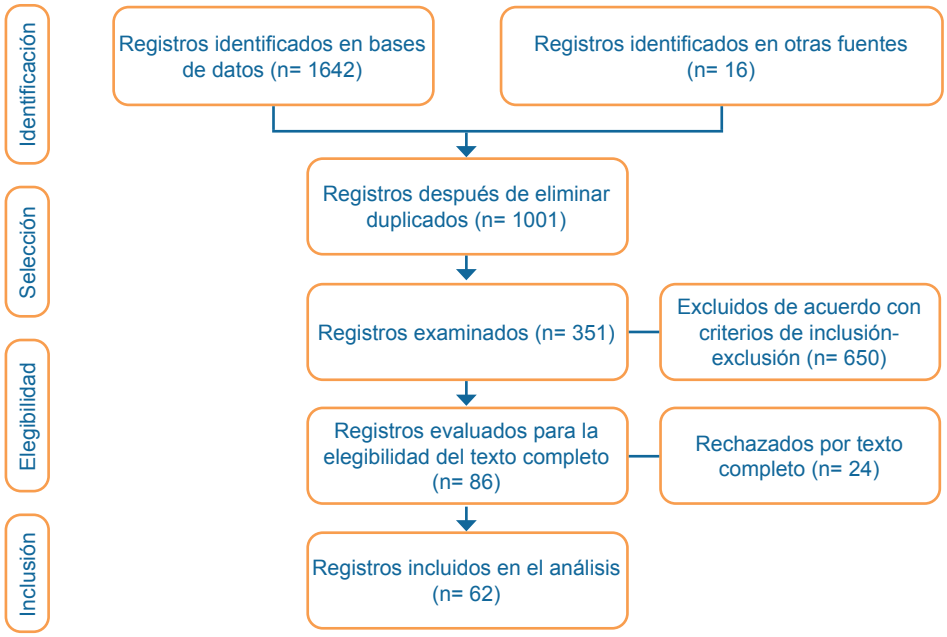
Se identificaron diferencias entre clinimetría y la psicometría, las cuales se apoyan en diversos aspectos vinculados con las características de las variables que se pretende medir, así como en el propósito de la medición. Son clásicas las obras de Feinstein[9] y Streiner[10] acerca de la clinimetría y numerosos autores han publicado libros y artículos acerca de la psicometría.[11-17] Una clara diferencia entre los procedimientos de la clinimetría y la psicometría se presentó en la publicación de DeVet y cols,[18] considerando que ambos enfoques representan utilidad significativa en el diseño y validación de escalas para su aplicación clínica (con múltiples dimensiones) o para la evaluación de constructos unidimensionales como en la metodología psicométrica.

Se han publicado al menos cuatro iniciativas que proponen estándares para evaluar las propiedades de medición; 1. Los atributos y criterios del Comité Asesor Científico del *Medical Outcomes Trust* (SAC-MOS);[19] 2. Los estándares de la Asociación Americana de Psicología (APA);[20] 3. Los criterios de calidad propuestos por Terwee y cols.,[21] y 4. La iniciativa COSMIN (*Consensus-based Standards for the selection of health Measurement Instruments*).[22,23]

La construcción y uso de escalas para evaluar variables o aspectos de interés cuyas características no permiten una medición directa, demanda desarrollar y aplicar el conocimiento de la metodología para lograr esta medición de una manera robusta.[4,5,24]

Con base en la revisión realizada, se organizaron los contenidos temáticos de este documento. La comprensión del propósito de estos elementos permitirá a los lectores dimensionar los alcances y

Figura 1. Proceso de selección de documentos



limitaciones de la medición que realicen con el instrumento o los instrumentos seleccionados.

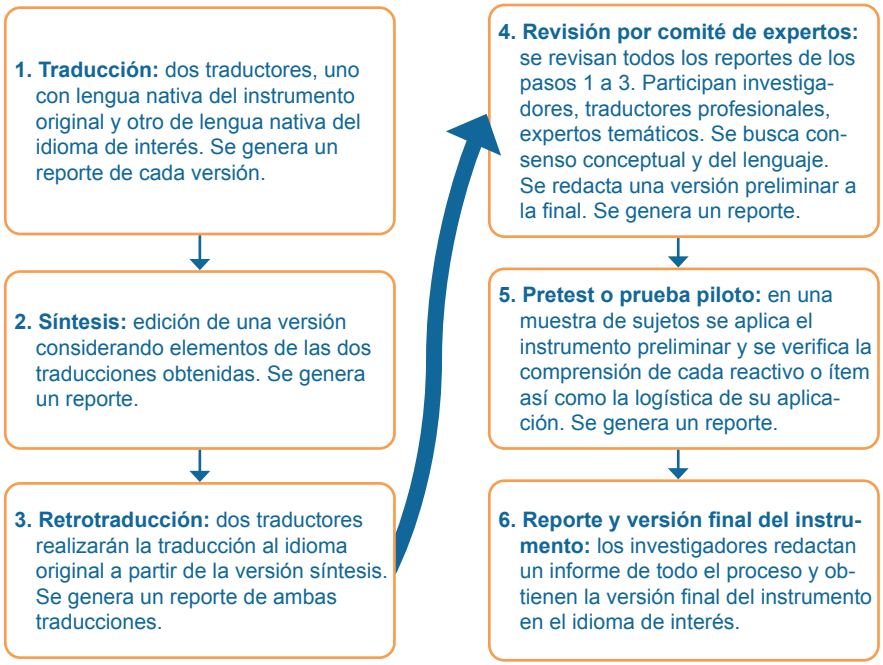
1. Traducción y adaptación cultural

La mayoría de los cuestionarios se han desarrollado en países de habla inglesa,[25] pero incluso dentro de estos países, los investigadores deben considerar a las poblaciones inmigrantes en los estudios de salud, especialmente cuando su exclusión podría conducir a un sesgo sistemático en los estudios sobre la utilización de los servicios sanitarios o la calidad de vida.

Conviene realizar esta búsqueda incluyendo artículos cuyos autores hayan desarrollado la metodología para la traducción y adaptación cultural del instrumento en cuestión.[26] La descripción de esta metodología se esquematiza en la figura 2.

Tras seleccionar los instrumentos que satisfagan las expectativas de medición del constructo o variables y completar su traducción y adaptación cultural, es necesario analizar su aplicación en distintas poblaciones, destacando sus propiedades psicométricas o clinimétricas. El término "constructo" en el ámbito de la investigación, se refiere a un concepto o idea abstracta que se analiza o estudia. Un constructo representa una entidad teórica que no es observable ni medible de una manera directa, como la inteligencia, la ansiedad, la funcionalidad familiar, la calidad de vida, el nivel de conocimientos, etc. Los constructos desempeñan un papel fundamental en el avance de disciplinas como la psicología, la sociología, la economía y la educación.

Figura 2. Metodología para traducción y adaptación cultural



Fuente: Modificado de Beaton y cols.[26]

2. Fiabilidad/confiabilidad

Los diversos documentos revisados, coinciden en la necesidad de evaluar la fiabilidad o confiabilidad de la medición que se obtenga mediante algún instrumento, independientemente de su denominación: cuestionario, índice, escala, test, inventario, etc. Cabe mencionar que a la fiabilidad también se le ha identificado como reproducibilidad.[23,27] Hay cuatro elementos que en conjunto contribuyen a la evaluación de la fiabilidad: consistencia interna, estabilidad, equivalencia y fiabilidad interobservador.[10, 28-32]

2.1 Consistencia interna

La consistencia interna de los ítems o reactivos de un instrumento o escala indica qué tan fuertemente se relaciona el contenido de los reactivos entre sí. Pueden utilizarse varias técnicas para medirla: a. La fiabilidad mitad y mitad, b. El alfa de Cronbach y c. La prueba de Kuder-Richardson. La más utilizada es el alfa de Cronbach;[32] esta técnica correlaciona cada reactivo con todos los demás que conforman el instrumento, se calcula la media de dichas correlaciones la cual se utiliza en la fórmula de Spearman-Brown para obtener su valor de probabilidad.[29]

Se considera el indicador idóneo porque da un único valor de consistencia y proporciona los datos de la técnica de la fiabilidad mitad y mitad. Los valores del alfa de Cronbach oscilan entre 0 y 1. Entre

más cercano es el valor a 1, denota una mayor consistencia interna. Según George y Mallery,[33] el valor alfa de Cronbach entre 0.7 y 0.8 puede considerarse aceptable, en el intervalo 0.8 y 0.9 es un nivel bueno y un valor superior a 0.9 es excelente; no obstante, Streiner considera que el valor máximo de alfa debe ser 0.9, valores mayores indican que hay redundancia o duplicación.[34] Para evitar esta redundancia, dichos ítems deben eliminarse; por lo general, se prefieren valores de alfa entre 0.80 y 0.90.

2.2 Estabilidad

La estabilidad alude a la consistencia de las respuestas obtenidas en repetidas ocasiones, por los mismos sujetos y en las mismas condiciones. La estimación de la estabilidad se obtiene con la técnica del test-retest. El método consiste en utilizar el instrumento en una misma muestra de sujetos en al menos dos ocasiones diferentes y comparar los resultados.

Se sugiere un intervalo de dos a cuatro semanas en aquellos instrumentos que midan variables estables.[12] Esta técnica tiene limitantes, como la posibilidad de que los sujetos en la segunda administración del cuestionario recuerden las respuestas de la primera, este efecto puede favorecer la obtención de un coeficiente de correlación erróneamente elevado. También es necesario tener en cuenta que, los sujetos en la segunda administración puedan contestar los ítems con menos atención que en la primera o no acepten que se les administre el cuestionario en dos ocasiones. Para valorar la estabilidad mediante test-retest se puede usar el coeficiente de correlación de Pearson para variables cuantitativas, el de Spearman-Brown para las nominales u ordinales o el coeficiente de correlación intraclase (CCI).[34]

2.3 Equivalencia

En la fase de diseño del instrumento, suele disponerse de dos o más versiones, también es posible encontrar dos o más versiones del mismo instrumento en un mismo idioma pero que ha sido aplicado en poblaciones diferentes; por ejemplo, en español aplicado en España y en español aplicado en Argentina o México. Para decidir cuál versión utilizar, se mide el grado de correlación entre las versiones aplicándolas sucesivamente a los sujetos en un mismo tiempo.

Valores de correlación por encima de 0.8 reflejan que los instrumentos son equivalentes. Si la correlación fuera inferior, no hay equivalencia y, por lo tanto, habrá que desarrollar otra versión para aplicarla en la población que se desea evaluar.

2.4 Fiabilidad inter-observador

La fiabilidad inter-observador o armonía interjueces, permite evaluar el grado de concordancia demostrado al repetirse una medición en condiciones idénticas a cargo de observadores distintos.

La ausencia de fiabilidad entre observadores puede surgir de las divergencias entre los instrumentos de medición o de la inestabilidad del atributo objeto de medición.[35,36] La fiabilidad inter-observador no es evaluable en instrumentos de auto aplicación, donde el mismo individuo es el que proporciona las respuestas, sin interferencia de los evaluadores en los resultados.

Se obtiene calculando el coeficiente de correlación de Pearson o Spearman. Se han descrito otras técnicas para evaluar la fiabilidad inter-observador, como los coeficientes de confiabilidad para escalas de calificación ordinales, la kappa de Cohen, el kappa ponderado linealmente y el kappa ponderado cuadráticamente; el coeficiente de correlación intraclase ICC (3,1), la correlación de Pearson, la rho de Spearman y la tau-b de Kendall.[37]

Los coeficientes kappa se utilizan comúnmente para cuantificar la confiabilidad en una escala categórica, mientras que los coeficientes de correlación se aplican comúnmente para evaluar la confiabilidad en una escala de intervalo.[38]

Estos coeficientes deben alcanzar puntuaciones por encima de 0.5 y se recomienda llegar a 0.7 para considerar que existe concordancia aceptable.[28,30] El diseño del instrumento, la elaboración de un manual específico y el entrenamiento de los observadores pueden ser elementos importantes para alcanzar valores de acuerdo con inter-observadores adecuados.

3. Validez

La validez se define como el grado en que una prueba o instrumento mide lo que pretende medir;[32] es un elemento esencial tanto en el diseño como en la comprobación de la utilidad de la medida realizada.

La validez se refiere al grado en que las puntuaciones reflejan el constructo subyacente previsto y se vincula con la interpretación de los resultados más que con el instrumento en sí. Lo más adecuado es considerarla como un argumento cuidadosamente estructurado en el que se reúne evidencia para apoyar o refutar las interpretaciones propuestas de los resultados.[8] Con fines didácticos y con fundamento en los procedimientos empleados, se puede clasificar la validez de un constructo en dos categorías, constructo teórico y constructo empírico. La evaluación de la validez incluye diferentes aspectos, los cuales son calificativos de este concepto y se esquematizan en la figura 3.

3.1 Validez de constructo teórico

Se refiere a los procedimientos que no incluyen todavía la aplicación del instrumento de medición a sujetos de estudio con el propósito de establecer una medición formal.

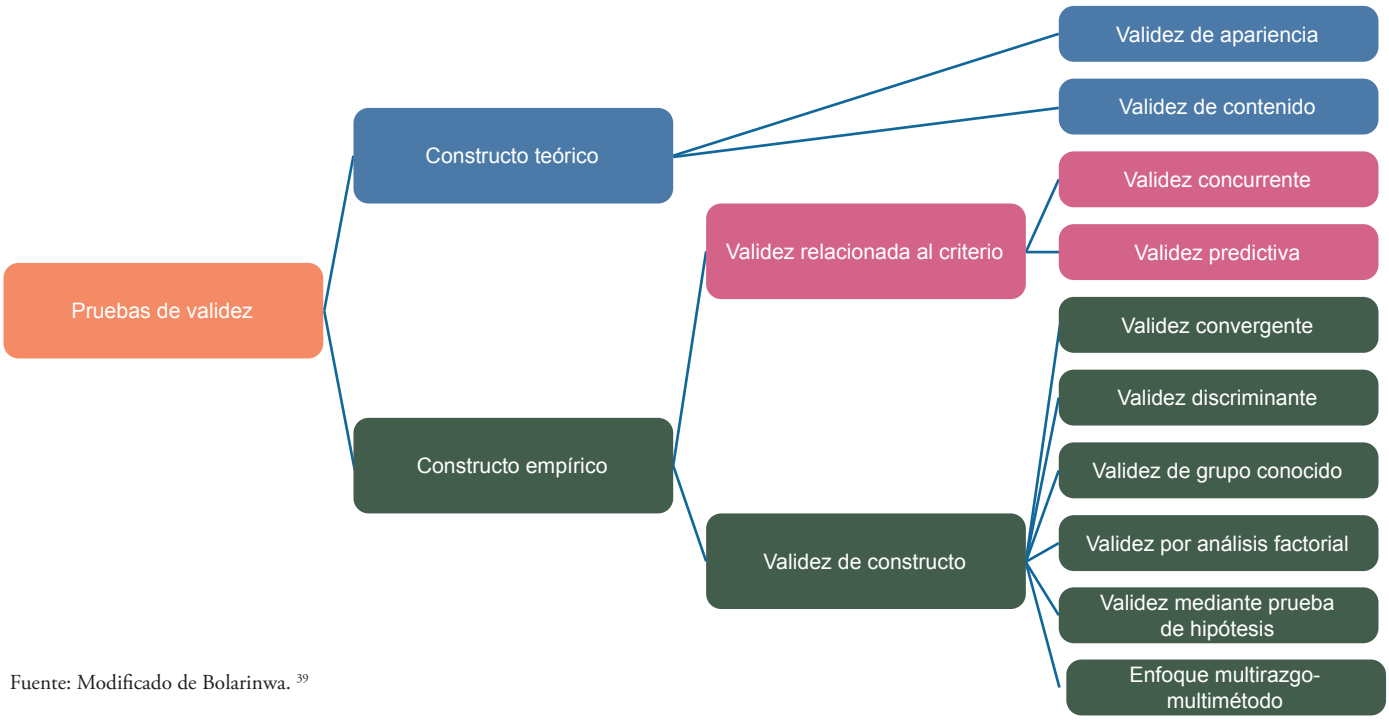
3.1.1 Validez de apariencia

Permite determinar la aceptabilidad que puede tener la escala en el escenario de aplicación.[5] Para obtenerla se deben conformar dos grupos, uno de expertos en el elemento temático central de lo que se busca medir mediante la escala o instrumento, quienes revisan y dictaminan si ésta realmente mide lo que se propone, y otro grupo que potencialmente sería incluido en un estudio para medir el constructo de interés, a este segundo grupo se le cuestiona para obtener su percepción de qué es lo que el instrumento mide. Cada grupo puede estar conformado por cuatro o cinco personas. La manera y criterios para seleccionar a los expertos, así como su número puede consultarse en el artículo de Escobar y Cuervo.[40]

3.1.2 Validez de contenido

La validez de apariencia y la de contenido están relacionadas. En la de apariencia se recurre a expertos y sujetos que respondan el instrumento del que se trate, en la validez de contenido se desarrolla un proceso para analizar la estructura del instrumento o la escala para garantizar que ésta, por medio de sus ítems, incluya todos los dominios de la entidad que se mide confirmando que el fenómeno objetivo de la medición está representado totalmente por sus ítems.[23,40,41]

Figura 3. Representación gráfica de los subtipos de las diversas formas de prueba de validez



Fuente: Modificado de Bolarinwa. [39]

Se han descrito al menos dos formas de medir la validez de contenido: a. Mediante el cálculo del índice de validez de contenido (CVI), y b. Mediante la aplicación de análisis factorial exploratorio-confirmatorio.[42-44]

El CVI a su vez, consta de dos mediciones, una vinculada con los reactivos y otra con la escala o cuestionario. En la tabla 1 se presentan los valores aceptables de CVI, los cuales difieren de acuerdo con el número de expertos que participen en la evaluación del instrumento.

La aplicación del análisis factorial exploratorio-confirmatorio se usa para obtener evidencias de las dimensiones subyacentes (componentes principales) que están presentes en el instrumento y que deberían corresponder en teoría, al constructo que se desea medir.

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, la adecuación de la muestra y los ítems evaluados, para lo cual se utilizan el test de esfericidad de Bartlett y el de Kaiser-Meyer-Olkin (KMO), este último se toma como satisfactorio cuando se obtienen valores mayores a 0.7.[3,43,45] 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 en valor absoluto. [45]

3.2 Validez de constructo empírico

La validez de constructo empírico incluye técnicas que requieren de la aplicación del instrumento a sujetos de estudio.

3.2.1 Validez de criterio

Es la capacidad de un cuestionario para medir qué tan bien una medida predice un resultado para otra medida.[41] En otras palabras, es el grado en que los puntajes obtenidos a partir de una escala son válidos, al compararlos con un estándar o patrón de referencia (criterio).[21,40] Con base en el momento en que se realice la comparación de resultados, es posible evaluar las dos características de esta propiedad: concurrencia y predictividad, de aquí los términos de validez concurrente y validez predictiva.[3]

3.2.1.1 Validez concurrente

En la validez concurrente, 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 (estándar de oro).

3.2.1.2 Validez predictiva

La validez predictiva es la capacidad de una prueba para medir algún evento o resultado en el futuro. Esto se evalúa mediante el coeficiente de correlación.[3] La validez predictiva se operacionaliza como el coeficiente de correlación nocional de Pearson entre el resultado de la herramienta de decisión y la medida relevante de utilidad clínica. Sin embargo, sería razonable operacionalizar la validez predictiva de otras maneras (por ejemplo, la correlación de rangos de Spearman o el área bajo la curva ROC).[46,47]

Para los instrumentos que miden los resultados informados por los pacientes relacionados con la salud (HRPRO) no existen estándares de oro, y se ha consensuado que la única excepción de un patrón oro es cuando se compara un instrumento acortado con la versión larga original.

Cuando se compara un nuevo instrumento con otro que no es el estándar de oro, se puede considerar como una validación de constructo, y se deben probar las hipótesis esperadas sobre la magnitud y dirección de la correlación entre los instrumentos.[21,23]

Tabla 1. Número de expertos y su implicación en el puntaje aceptable de corte del índice de validez de contenido

Número de expertos	Valores aceptables de CIV
Dos	0.80
Tres a cinco	1.00
Seis a ocho	0.83
Nueve	0.78

Modificado de Yusoff. [43]

3.2.2 Validez de constructo

Es la adherencia de una evaluación al conocimiento y teoría existente sobre el concepto que se está midiendo.[48] Si un cuestionario carece de validez de constructo, será difícil interpretar sus resultados y no se podrán obtener inferencias acerca de un dominio de conducta.[42] Se define como el grado en que un instrumento mide el rasgo o constructo teórico que pretende medir.[49-51] No tiene un criterio de comparación, sino que utiliza una construcción hipotética para la comparación.[50,51] Es la medida de validez más valiosa y difícil de fundamentar. Básicamente, es una medida de cuán significativa es la escala o instrumento cuando se utiliza en la práctica.[52]

Para lograr la validez de constructo de un instrumento, se debe considerar la aplicación de al menos dos de los cinco procedimientos que se describen enseguida.

3.2.2.1 Validez convergente y discriminante

La validez convergente es una correspondencia o convergencia entre constructos que son teóricamente similares. Los coeficientes de correlación entre ítems serían altos en un instrumento que tiene validez convergente. Por el contrario, la validez discriminante es la capacidad del instrumento para diferenciar o discriminar entre constructos que son teóricamente diferentes. Los coeficientes de correlación entre ítems serían bajos en un instrumento con validez discriminante.

La validez convergente se apoyaría si las puntuaciones fueran similares. La validez discriminante estaría respaldada si las correlaciones fueran bajas entre la herramienta del dolor y una medida de comodidad o bienestar.[49]

Se recomiendan valores de validez convergente por encima de r= 0.70, mientras que deben evitarse aquellos por debajo de r= 0.50. Es muy conveniente que los investigadores desarrollen y comuniquen los datos de la validez convergente.[53]

3.2.2.2 Validez de grupo conocido

Existen diversas formas de evaluar la validez de constructo de un instrumento: validez de grupo conocido, análisis factorial, pruebas de hipótesis, y el enfoque multirasgo-multimétodo (MT-MM).[50]

En el enfoque de grupo conocido, se toman muestras de dos grupos que se sabe que son altos y bajos en el constructo que se está midiendo. Las puntuaciones medias de los dos grupos deberían diferir significativamente en la dirección esperada si el instrumento es válido.

3.2.2.3 Validez por análisis factorial

Es una extensión empírica de la validez de contenido, esto se debe a que valida el contenido del constructo en el que se emplea un modelo estadístico llamado análisis factorial, el cual se utiliza cuando el constructo de interés se encuentra en muchas dimensiones que forman diferentes dominios de un atributo general.[54] Un factor es una combinación de elementos de prueba los cuales se agrupan y definen una parte del constructo.

Los elementos no relacionados no definen el constructo y deben eliminarse del instrumento evaluado.[55] El análisis factorial exploratorio (EFA) ayuda a los investigadores a identificar los diversos factores que definen el constructo, este análisis se utiliza para identificar la mayor varianza en las puntuaciones con el menor número de factores, expresada estadísticamente como un eigenvalue >1,0.

El análisis factorial confirmatorio (AFC) generalmente sigue al EFA e incluye conocimientos teóricos para probar más a fondo la validez de constructo de un instrumento. Los expertos en psicometría no están de acuerdo con el número de participantes necesarios para el análisis factorial y debería haber al menos cinco (algunos autores dicen 20) veces más encuestados en la muestra que variables a analizar.[43,55,56]

3.2.2.4 Validez mediante prueba de hipótesis

La prueba de hipótesis se basa en un marco teórico e indica la dirección esperada de las puntuaciones de la medida. La validez de constructo se respalda si las puntuaciones reflejan el marco tal como lo plantea la hipótesis. Suponga que un investigador desarrolla un instrumento diseñado para medir el agotamiento en médicos residentes. El investigador plantea la hipótesis de que los residentes que trabajan en servicios con pacientes con alta gravedad tienen mayor agotamiento que los residentes en servicios con pacientes estables. Para probar la hipótesis, el instrumento se administra a una muestra de residentes que trabajan en una unidad de cuidados intensivos (UCI) y a una muestra de los que trabajan en el área de hospitalización general. Un mayor nivel de agotamiento entre los residentes de la UCI que entre los residentes del área de hospitalización general podría considerarse evidencia de la validez de constructo del instrumento recientemente desarrollado.

3.2.2.5 Enfoque multirasgo-multimétodo (MT-MM)

La matriz multirrasgo-multimétodo (MT-MM) permite evaluar la validez de constructo, en particular la validez convergente y la discriminante. El investigador debe medir un conjunto de rasgos utilizando múltiples métodos en un estudio.

Con medidas de diferentes rasgos mediante diferentes métodos, los investigadores pueden generar una matriz de correlación: la matriz multirasgo-multimétodo. Los datos de correlación en la matriz se utilizan para juzgar la calidad de la medición de los constructos previstos.[57]

El método MT-MM se puede utilizar siempre que se midan dos o más constructos con dos o más metodologías. Diferentes medidas del mismo constructo deberían correlacionarse altamente entre sí (converger) y diferentes constructos deberían mostrar una baja correlación entre sí (discriminar).[49]

4. Capacidad de respuesta

El grupo de estándares basados en consenso para la selección de instrumentos de medición de salud (COSMIN), la define como "la capacidad de un instrumento para detectar cambios a lo largo del tiempo en el constructo a medir".[58] Según la taxonomía COSMIN, la validez se enfoca en una puntuación única, mientras que la capacidad de respuesta se refiere a la validez de una puntuación de cambio.[59]

La capacidad de respuesta sólo es relevante para los instrumentos de medición que se utilizan en aplicaciones evaluativas, es decir, cuando el instrumento se utiliza en un estudio longitudinal para medir el cambio a lo largo del tiempo.[60] La capacidad de respuesta se enfoca en la utilidad de un cuestionario para detectar cambios clínicamente importantes a lo largo del tiempo.[61]

Se debe tener cuidado al evaluar la capacidad de respuesta de un instrumento de medición dentro del mismo estudio en el que se emplea como medida de resultado, ya que esto impide diferenciar entre la calidad del instrumento y el efecto de la intervención.[62]

Discusión

La selección de instrumentos para la medición de variables en las disciplinas de la salud requiere de un enfoque crítico que se base en una exhaustiva revisión de la literatura científica. Esta revisión debe:

1. Permitir el análisis de las características descritas en el presente artículo
2. Enfocarse en fuentes primarias, es decir, en artículos originales publicados por los autores responsables del diseño del instrumento de medición, así como en artículos cuyo contenido exprese en forma adecuada y completa, las propiedades de fiabilidad, validez y en su caso, de capacidad de respuesta

Esta revisión permitió el análisis de artículos vinculados con la validez como concepto y el proceso de validación de instrumentos de utilidad y pertinencia en aéreas como la educación, la psicología, las ciencias sociales y de la salud.[5,22,23,45,52,63-68]

En contraste, son escasas las publicaciones acerca de la descripción y sistematización del proceso que implica la validación completa, rigurosa y exhaustiva de estas herramientas de medida. Esto permite afirmar que, si bien hay interés por investigar y describir estos aspectos metodológicos desde un enfoque teórico fundamentado, principalmente desde las perspectivas psicométrica y clinimétrica, se ha carecido de elementos suficientes para consensuar aspectos fundamentales como la terminología, conceptos y definiciones, procedimientos estadísticos, criterios psicométricos mínimos para evaluar y, como consecuencia, los procesos de construcción, adaptación y validación de las mediciones y sus instrumentos en el área de la salud.[21]

En este documento, se identificaron artículos con ejemplos de instrumentos de medición que han sido sujetos a diversos procedimientos para respaldar su fiabilidad y validez y que se utilizan en la investigación y en el ámbito clínico, pero también se advierten deficiencias en los procedimientos de construcción y adaptación al aplicar las pruebas psicométricas en forma incompleta o inadecuada.

La validez de constructo sigue estando subestimada o pocas veces reportada con todos los procedimientos, sus valores y su significado.[49] La mayoría de los investigadores informan sobre la confiabilidad de la consistencia interna (Alpha de Cronbach), pero pocos artículos indican algún tipo de prueba de evaluación de constructo. Cuando se afirmaba la confiabilidad de la nueva prueba, con frecuencia no se incluían los intervalos de tiempo ni las correlaciones. La solidez de las pruebas psicométricas depende de muchos factores importantes, como la teoría subyacente, las consecuencias de medidas débiles y las implicaciones de los hallazgos.

Contrariamente a la práctica común, la evidencia sobre la idoneidad de una herramienta de medición exige más de uno o dos estudios sobre su estructura dimensional o la magnitud de las cargas factoriales. En una revisión sistemática de la literatura, Hamilton y col.,[69] evidencian la conveniencia de revisar las propiedades psicométricas de diversas herramientas para la evaluación familiar, varias de ellas, con más de diez años de haberse diseñado y con un gran número de publicaciones que en forma reiterada analizan sus propiedades psicométricas.

El desarrollo y perfeccionamiento de diferentes versiones de un instrumento también son vitales, de modo que las investigaciones realizadas en distintas poblaciones mantengan la comparabilidad entre sí. El proceso de adaptación

transcultural es tan complejo como el desarrollo de un nuevo instrumento. Para contar con evidencia suficiente acerca de la idoneidad de una herramienta de medición se requieren más de uno o dos estudios sobre su estructura dimensional o la magnitud de las cargas factoriales.[63]

Una pregunta frecuente es si es necesario completar todas las fases y etapas para considerar un instrumento adecuado para la investigación o para su uso dentro de los servicios de salud. Esta pregunta no tiene una respuesta fácil, pero al menos dos criterios pueden servir de guía:

a. La adecuada planificación y desarrollo de los procedimientos para la elaboración del constructo y sus dimensiones favorece la fiabilidad y la validez en sus distintos enfoques

b. El rigor en el diseño mejora las propiedades psicométricas y optimiza la eficiencia, al prevenir o resolver problemas desde el inicio. Por ello, la revisión detallada de los procedimientos que originaron el instrumento es fundamental en la búsqueda de literatura

Conclusiones

Para seleccionar adecuadamente un instrumento de medición de variables cuyas características no permiten una medición directa, es determinante la búsqueda y evaluación de las propiedades descritas en los instrumentos de medición. Al satisfacer adecuadamente las pruebas de fiabilidad y validez, es posible legitimar comparaciones entre diversas poblaciones, obtener resultados válidos 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.

Contribución de los autores

FJFG-C: conceptualización, búsqueda y selección de artículos, análisis de la información, redacción del artículo, revisión del contenido. GL-O: organización y jerarquización de los temas y subtemas, redacción y revisión del contenido.

Todos los autores aprueban la publicación del presente escrito.

Financiamiento

La presente investigación no recibió financiamiento alguno.

Conflictos de interés

Los autores declaran no tener conflictos de interés.

Referencias

1. Luján-Tangarife JA, Cardona-Arias JA. Construcción y validación de escalas de medición en salud: revisión de propiedades psicométricas. Arch Med. 2015;11(3):1-10.
2. Manzano Patiño AP. Introducción a los modelos de ecuaciones estructurales. Investig Educ Médica. 2018;7(25):67-72.
3. Batista-Foguet JM, Coenders G, Alonso J. Análisis factorial confirmatorio. Su utilidad en la validación de cuestionarios relacionados con la salud. Med Clin (Barc). 2004;122(Supl 1):21-27.
4. Sánchez PR, Gómez RC. Conceptos básicos sobre la validación de escalas. Rev Col Psiquiatr. 1998;27(2):121-130.
5. Sánchez R, Echeverry J. Validación de escalas de medición en salud. Rev Salud Pública (Bogotá). 2004;6(3):302-318.
6. Lara M, Ortega H. ¿Clinimetría o Psicometría? Medición en la práctica psiquiátrica. Salud Ment. 1995;18(4):33-40.
7. Beckman TJ, Ghosh AK, Cook DA, et al. How reliable are assessments of clinical teaching? J Gen Intern Med. 2004;19:971-977.
8. Ramada-Rodilla JM, Serra-Pujadas C, Delclós-Clanchet GL. Adaptación cultural y validación de cuestionarios de salud: revisión y recomendaciones metodológicas. Salud Publica Mex. 2013;55(1):57-66.
9. Feinstein AR. Clinimetric perspectives. J Chronic Dis. 1987;40(6):635-640.
10. Streiner DL, Norman GR. Health measurement scales: A practical guide to their development and use. 5th ed. Oxford: Oxford University Press; 1995.
11. Nunnally JC, Bernstein IJ. Teoría psicométrica. México: McGraw-Hill; 1995.
12. Nunnally JC. Introducción a la medición psicológica. Buenos Aires: Paidós; 1977.
13. Meneses J, Barrios M, Bonillo A, Cosculluela A, Lozano LM, Turbani J, Valero S. Psicometría. Barcelona: Editorial UOC; 2013.
14. Ramos VZ. Psicometría básica. Bogotá: Fundación Universitaria del Área Andina; 2018.
15. González LLFM. Instrumentos de evaluación psicológica. La Habana: Editorial Ciencias Médicas; 2007.
16. Cortada De K. Importancia de la investigación psicométrica. Rev Lat Psic. 2002;34(3):229-240.
17. Aragón B. Fundamentos psicométricos en la evaluación psicológica. Rev Elect Psic Iztacala. 2004;7(4):23-43.
18. Devet H, Terwee C, Bouter L. Clinimetrics and psychometrics: two sides of the same coin. J Clin Epidemiol. 2003;56(12):1146-1147.
19. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res. 2002;11:193-205.
20. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 1999.
21. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34-42.
22. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, 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(4):539-549.
23. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737-745.
24. Lamprea M, Gómez-Restrepo C. Validez en la evaluación de escalas. Rev Colom Psiquiat. 2007;36(2):340-348.
25. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol. 1993;46:1417-1432.
26. 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(24):3186-3191.
27. Manterola C, Grande L, Otzen T, García N, Salazar P, Quiroz G. Confiabilidad, precisión o reproducibilidad de las mediciones. Métodos de valoración, utilidad y aplicaciones en la práctica clínica. Rev Chilena Infectol. 2018;35(6):680-688.
28. Kerlinger FN, Lee HB. Investigación del comportamiento. Métodos de investigación en ciencias

sociales. 4ª ed. México: McGraw-Hill/Interamericana Editores; 2002. p. 581-602.
29. Streiner DL, Norman GL, Cairney J. Health measurement scales: A practical guide to their development and use. UK: Oxford University Press; 2015.
30. Hurley WL, Denegar CR, Hertel J. Métodos de investigación: Fundamentos de una práctica clínica basada en la evidencia. Barcelona: Wolkers Kluwer Health Lippincott Williams & Wilkins; 2012. p. 139-54.
31. Mcdowell I. Measuring health: A guide to rating scales and questionnaires. New York: Oxford University Press; 2006. p. 39-45.
32. Cascaes da Silva F, Gonçalves E, Valdivia Arancibia BA, Graziele Bento S, Da Silva Castro TL, Soleman Hernandez SS, et al. Estimadores de consistencia interna en las investigaciones en salud: el uso del coeficiente alfa. Rev Peru Med Exp Salud Publica. 2015;32(1):129-138.
33. George D, Mallery P. IBM SPSS Statistics 25 Step by step: A simple guide and reference. New York: Routledge/Taylor & Francis; 2019.
34. Streiner DL. Starting at the beginning: An introduction to coefficient alpha and internal consistency. J Pers Assess. 2003;80(1):99-103.
35. Évaluation des technologies de la santé [Internet] [citado 2024 mar 19]. Disponible en: https://htaglossary.net/fiabilidad-entre-observadores-(n.f.)
36. de Raadt A, Warrens MJ, Bosker RJ, Kiers HAL. A comparison of reliability coefficients for ordinal rating scales. J Classif. 2021;38(3):519-543.
37. Jacobsson U, Westergren A. Statistical methods for assessing agreement for ordinal data. Scand J Caring Sci. 2005;19:427-31.
38. Carvajal A, Centeno C, Watson R, Martínez M, Sanz Rubiales Á. ¿Cómo validar un instrumento de medida de la salud? An Sist Sanit Navar. 2011;34(1):63-72.
39. Bolarinwa O. Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. Niger Postgrad Med J. 2015;22(4):195.
40. Escobar-Pérez J, Cuervo-Martínez A. Validez de contenido y juicio de expertos: una aproximación a su utilización. Avances Medición. 2008;6:27-36.
41. García de YPMA, Rodríguez SF, Carmona OL. Validación de cuestionarios. Reumatol Clin. 2009;5:171-177.
42. Tsang S, Royse C, Terkawi A. Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. Saudi J Anaesth. 2017;11(5):80-89.
43. Yusoff MSB. ABC of content validation and content validity index calculation. Educ Med J. 2019;11(2):49-54.
44. Pérez-Gil JA, Chacón MS, Moreno RF. Validez de constructo: el uso del análisis factorial exploratorio-confirmatorio para obtener evidencias de validez. Psicothema. 2000;12(2):442-446.
45. Montero RE. Referentes conceptuales y metodológicos sobre la noción moderna de validez de instrumentos de medición: implicaciones para el caso de personas con necesidades educativas especiales. Actual Psicol. 2013;27(114):113-128.
46. Soto J. Validez predictiva de los cuestionarios: ¿qué es y por qué es importante su conocimiento. Reumatol Clin. 2010;6(1):1-4.
47. Scannell JW, Bosley J, Hickman JA, Dawson GR, Truebel H, Ferreira GS, et al. Predictive validity in drug discovery: what it is, why it matters and how to improve it. Nat Rev Drug Discov. 2022;21(12):915-931.
48. Ahmed I, Ishtiaq S. Reliability and validity: Importance in Medical Research. J Pak Med Assoc. 2021;71(10):2401-2406.
49. DeVon HA, Block ME, Moyle-Wright P, Ernst DM, Hayden SJ, Lazzara DJ, et al. A psychometric toolbox for testing validity and reliability. J Nurs Scholarsh. 2007;39(2):155-164.
50. Strauss ME, Smith GT. Construct validity: advances in theory and methodology. Annu Rev Clin Psychol. 2009;5(1):1-25.
51. Anderson JL, Sellbom M. Construct validity of the DSM-5 Section III personality trait profile for borderline personality disorder. J Pers Assess. 2015;97(5):478-486.
52. Drost EA. Validity and reliability in social science research. Educ Res Perspect. 2011;38:105-123.
53. Carlson KD, Herdman AO. Understanding the impact of convergent validity on research results. Organ Res Methods. 2012;15(1):17-32.
54. Douglas H, Bore M, Munro D. Construct validity of a two-factor model of psychopathy. Psychology (Irvine). 2012;3(3):243-248.
55. Mcdowell I. Measuring Health: A guide to rating scales and questionnaires. New York: Oxford University Press; 2006. p. 53-4.
56. Dhillon HK, Zaini MZA, Quek KF, Singh HJ, Kaur G, Rusli BN. Exploratory and confirmatory factor analyses for testing validity and reliability of the Malay language questionnaire for urinary incontinence diagnosis (QUID). Open J Prev Med. 2014;4(11):844-851.
57. Shen F. Multitrait-Multimethod Matrix. The International Encyclopedia of Communication Research Methods. Wiley; 2017. p. 1-6.
58. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL. International consensus on taxonomy, terminology, and definitions of measurement properties: Results of the COSMIN study. J Clin Epidemiol. 2010;63:737-745.
59. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Med Res Methodol. 2010;10:22.
60. Terwee CB. Responsiveness to Change. En: Encyclopedia of Quality of Life and Well-Being Research. Dordrecht: Springer Netherlands; 2014. p. 5547-50.
61. Kievit AJ, Kuijer PPFM, Kievit RA, Sierevelt IN, Blankevoort L, Frings-Dresen MHW. A reliable, valid and responsive questionnaire to score the impact of knee complaints on work following total knee arthroplasty: The WORQ. J Arthroplasty. 2014;29(6):1169-1175.e2.
62. De Vet H, Terwee CB, Mokkink LB, Knol DJ. Measurement in medicine. A practical guide. Cambridge: Cambridge University Press; 2011.
63. Reichenheim M, Bastos JL. What, what for and how? Developing measurement instruments in epidemiology. Rev Saude Publica. 2021;55:40.
64. Berry JW, Poortinga YH, Segall MH, Dasen PR. Cross-cultural psychology: research and applications. New York: Cambridge University Press; 2002.
65. Herdman M, Fox-Rushby J, Badia X. "Equivalence" and the translation and adaptation of health-related quality of life questionnaires. Qual Life Res. 1997;6(3):237-247.
66. Gómez-Clavelina FJ, Irigoyen-Coria AE, Ponce-Rosas ER. Selección y análisis de instrumentos para la evaluación de la estructura y funcionalidad familiar. Arch Med Fam. 1999;1(2):45-57.
67. Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. Am J Health Syst Pharm. 2008;65(23):2276-2284.
68. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34-42.
69. Hamilton E, Carr A. Systematic review of self-report family assessment measures. Fam Process. 2016;55(1):16-30.