Artículo especial

Validación de instrumentos clínicos: aspectos esenciales

Title: Clinical Instruments Validation: Key Aspects

Haydee Alejandra Martini-Blanquel[*]

Resumen

Este estudio explora algunos aspectos clave sobre la validación de instrumentos utilizados en la recolección de datos clínicos, lo cual es parte de un proceso crítico para el diagnóstico y tratamiento en el campo médico. La clinimetría subraya la importancia de la validez y confiabilidad en el uso de instrumentos para la identificación y medición precisa de signos y síntomas. La validez asegura que el instrumento mida efectivamente la variable de interés, mientras que la confiabilidad indica que el uso repetido del mismo instrumento producirá resultados consistentes. El artículo aborda aspectos relacionados con la construcción y validación de instrumentos, incluyendo la fundamentación teórica y empírica, la validación por jueces expertos, pruebas de premuestreo, así como evaluaciones de constructo y criterio. Además, se aborda la aplicación del análisis factorial exploratorio y confirmatorio para validar la estructura interna de los instrumentos. Este trabajo resalta la importancia de la precisión en la práctica clínica y ofrece un marco detallado para garantizar la eficacia y relevancia de los instrumentos en diferentes contextos poblacionales.

Palabras clave: estudios de validación, fiabilidad, validez.

Recibido: 06/11/2023
Aceptado: 18/04/2024

[*]Unidad de Medicina Familiar No. 33, Instituto Mexicano del Seguro Social.

Correspondencia:
Haydee Alejandra Martini-Blanquel
haydee.martini@imss.gob.mx

Sugerencia de citación: Martini-Blanquel HA. Validación de instrumentos clínicos: aspectos esenciales. Aten Fam. 2024;31(3): 185-192. http://dx.doi.org/10.22201/fm.14058871p.2024.388840

Summary

This study explores some key aspects of the validation of instruments used in clinical data collection, which is part of a critical process for diagnosis and treatment in the medical field. Clinimetrics underscores the importance of validity, and reliability in the use of instruments for the accurate identification, and measurement of signs and symptoms. Validity ensures that the instrument effectively measures the variable of interest, while reliability indicates that repeated use of the same instrument will produce consistent results. The article addresses aspects related to instrument construction and validation, including theoretical and empirical foundations, validation by expert judges, pre-sampling tests, as well as construct and criterion evaluations. In addition, the application of exploratory and confirmatory factor analysis to validate the internal structure of the instruments is addressed. This work highlights the importance of accuracy in clinical practice and provides a detailed framework for ensuring the efficacy and relevance of instruments in different population contexts.

Key Words: Validation Studies; Reliabiliy; Validity.

Introducción

La recopilación de los datos clínicos es un proceso de gran valor para el médico, pues permite establecer diagnósticos oportunos, tratamientos y pronósticos en los pacientes que se atienden. Sin embargo, en el caso de muchas entidades nosológicas que son objeto de una investigación —adicional a la recolección de datos en una historia clínica o una nota médica —, se requiere explicar y cuantificar de forma más exacta signos y síntomas. Para lograrlo, la metodología utilizada debe ser la correcta, ya que ello influirá en la toma de decisiones, tanto para concretar un diagnóstico como para prescribir medicamentos o establecer el pronóstico de alguna enfermedad.[1] De ahí que los conceptos abstractos (signos y síntomas) se convierten en datos científicos que se pueden medir, a esto se le conoce con el término de "clinimetría".[2]

De tal forma, el término "medición" siempre está presente en medicina, como en otras áreas de la salud. Por ejemplo, cuando se sospecha que una persona tiene sobrepeso, el médico calculará el índice de masa corporal (IMC) y posteriormente, utilizará la escala de la Organización Mundial de la Salud para poder confirmar su sospecha diagnóstica.[3] Asimismo, cuando desea corroborar el diagnóstico de trastorno depresivo, es probable que aplique alguna escala aceptada y reconocida; en este caso, podría ser el inventario de Beck[4] o la escala de valoración de Hamilton.[5] Sin embargo, a pesar de la cotidianidad con que son utilizadas estas herramientas, es frecuente que se desconozca su importancia, el proceso para elaborarlas y cómo ayudan al quehacer diario del médico.

Por lo descrito anteriormente, es importante que todo instrumento utilizado en la clínica represente de forma adecuada los conceptos o variables que quiere medir el médico con la mayor exactitud posible; a este concepto se le conoce como validez. Así, cuando se miden de esta forma las variables de interés, es más sencillo dar respuestas cercanas a la realidad.[1]

Confiabilidad y validez de un instrumento

La validez hace referencia al grado en que un instrumento mide realmente la variable que pretende medir.[6] Dicha validez está directamente relacionada con la confiabilidad, que tiene que ver con que un fenómeno, cuando es medido muchas veces con el mismo instrumento, otorga los mismos resultados. [1]

Esta definición de validez es la que encontramos de forma habitual en la literatura; sin embargo, desde hace décadas, existen otras propuestas, en las que se señala que un instrumento es válido en tanto sus resultados ayuden a realizar inferencias e interpretaciones y por lo tanto, existan consecuencias sociales y éticas de su aplicación.[7] Por ejemplo, cuando un test nos permite elegir qué pacientes son candidatos a recibir un tratamiento farmacológico y por ende, mejorar su estado de salud, o cuando una escala identifica a personas con riesgo elevado de padecer cierta enfermedad y con ello, el médico implementa medidas de prevención.

Sin embargo, el hecho de que un instrumento sea confiable no garantiza que sea exacto en sus mediciones. Lo anterior aplica porque cualquier instrumento, al ser elaborado en un determinado contexto, puede no ser útil en otra población; inclusive, la confiabilidad suele modificarse cuando el instrumento se adapta a un determinado tipo de personas o se ajusta a un idioma diferente.

De tal forma, el proceso de validación es permanente, lo cual implica realizar de forma constante comprobaciones que nos muestren que un instrumento es adecuado en un grado aceptable, considerando siempre los objetivos para los que fue creado y la población a la cual está dirigido.

Proceso para la construcción y validación de instrumentos de investigación

Tomando en cuenta lo anterior, los pasos básicos para llevar a cabo la construcción y validación de un instrumento son cuatro: 1. Fundamentación teórica y empírica del instrumento (considerando los objetivos para los que será creado), 2. Validación del instrumento por los jueces, 3. Prueba premuestreo, 4. Validación de constructo y de criterio y 5. Cálculo de la confiabilidad (consistencia interna) del instrumento.

1. Fundamentación teórica y empírica del instrumento

Los instrumentos de medición en la investigación clínica, desde un enfoque cuantivativo, deben ser elaborados considerando varios conceptos y criterios; concepto es sinónimo de constructo, el cual se desarrolla con la finalidad de lograr una medición con rigor científico. De esta forma, todo instrumento debe tener una base teórica y una empírica.[8]

a. Fundamento teórico

Se relaciona con los conceptos no observables de forma directa. Por ejemplo, si se quiere construir un instrumento que mida la variable "calidad de vida relacionada con la salud en pacientes con artritis reumatoide", tendremos que revisar la literatura y determinar si efectivamente, ya existen herramientas que midan este concepto. Después de una revisión detallada, el investigador decidirá si la información existente es cercana o no a aquello que se pretende medir (calidad de vida relacionada con la salud en pacientes con artritis reumatoide). En este punto, es posible que el concepto se encuentre parcialmente definido o que no haya antecedente de él en la literatura; de ser así, será necesario complementar o elaborar en su totalidad una propuesta teórica, seleccionando los conceptos y posteriormente, los indicadores, que son aquellas manifestaciones externas que permiten medir un constructo o un concepto.[9]

Aquí, el investigador se convierte en el experto número uno del tema elegido y por lo tanto, tiene la capacidad de generar nuevos conceptos.[10] Sin embargo, otras opciones para obtener información y elaborar la propuesta teórica son: 1) solicitar apoyo de jueces o expertos en el tema o 2) realizar entrevistas a una población similar a la que estará dirigido el instrumento (sobre todo en caso de aquellos dirigidos a pacientes).

Al momento de realizar la búsqueda, es posible que el constructo que intentamos medir tenga un solo atributo, en cuyo caso será de tipo unidimensional. No obstante, muchos de los instrumentos utilizados en la práctica clínica provienen de constructos complejos, como es el caso del ejemplo antes mencionado (calidad de vida relacionada con la salud), pues si revisamos la literatura encontraremos que para su evaluación se consideran diversos aspectos como la movilidad, el bienestar emocional, el cuidado personal, etcétera.[11] Por ello, estos instrumentos se denominan multidimensionales, pues consideran varios indicadores para medir una sola característica. Es así como el investigador principal seleccionará los conceptos e indicadores que quiere integrar en su instrumento y los resumirá lo más posible, evitando que se repitan.[10]

Finalmente, es importante resaltar que en este momento no es necesario cuestionarnos a profundidad sobre la pertinencia de los conceptos incluidos, pues más adelante se llevarán a cabo procedimientos específicos para comprobarlo.

b. Fundamento empírico

La parte empírica hace referencia a la adecuación del instrumento con base en la teoría, es decir, en conceptos ya definidos; si dicha teoría se encuentra bien fundamentada y es congruente, otorgará cierta facilidad para la redacción de los ítems o preguntas.

Siguiendo el ejemplo anterior sobre "calidad de vida relacionada con la salud en pacientes con artritis reumatoide", después de elaborar nuestra propuesta teórica o complementarla, debemos asegurarnos de que exista congruencia entre el objetivo del instrumento, los conceptos y los ítems que vamos a incluir. Entonces, si vamos a medir "calidad de vida relacionada con la salud en pacientes con artritis reumatoide", no sería lógico que elaboremos reactivos sobre "calidad de vida" únicamente, pues conceptualmente son términos diferentes.

Posteriormente, necesitamos precisar las características de los ítems, es decir, si serán enunciados en forma de aseveración o preguntas de opción múltiple, así como el tipo de respuestas, etcétera. En la construcción de los ítems, se deben tener presentes el listado de tópicos o conceptos que se eligieron en la fase anterior (fundamento teórico), así como las características de la población a quien está dirigido el instrumento (personal médico, niños, adultos mayores, mujeres, personas con alguna discapacidad, etcétera). Una vez que se tengan los ítems y el listado de las posibles opciones de respuesta, se establece la escala de medición y el tipo de instrumento, lo cual dependerá de la naturaleza del concepto y los indicadores (atributos) a evaluar. Por ejemplo, si

deseamos evaluar el consumo de ciertos alimentos, quizá lo más conveniente sea un cuestionario con respuesta de opción múltiple; por otro lado, si lo que queremos es medir el nivel de satisfacción de los pacientes en la atención brindada, entonces se pueden elaborar enunciados en forma de aseveración, con respuestas tipo Likert, en el que el encuestado debe indicar su acuerdo o desacuerdo en cada aseveración, cuyo puntaje mínimo se otorga cuando hay mayor desacuerdo y el máximo cuando hay mayor acuerdo. Generalmente, las opciones de respuesta numéricas van del 1 a 5.[12]

Hasta este momento, todavía no se tiene un instrumento estructurado; aún hay oportunidad de formular más preguntas y de cambiar la forma y el orden de estas, además de los conceptos previamente obtenidos. Lo anterior quiere decir que esta es una fase flexible, en la que podemos adicionar elementos no contemplados previamente, sin perder de vista el concepto y los indicadores que queremos evaluar.

En resumen, si existe una relación fuerte entre los indicadores, que son las respuestas observadas y los conceptos no observables, podemos decir que existe aplicabilidad empírica de las proposiciones teóricas y, por lo tanto, realizaremos una medición adecuada, cuyos resultados nos lleven a un mejor entendimiento del fenómeno estudiado.[7]

2. Validación del instrumento por los jueces

El investigador principal selecciona a los jueces, ante los cuales presentará el instrumento en su versión inicial. Los jueces deben estar familiarizados con la investigación y especialmente, con el proceso de validación de instrumentos, pero no necesariamente están relacionados con el objeto de estudio en cuestión. Por ello, un juez no hará aportaciones al contenido del instrumento; su tarea principal será evaluar los ítems que se construyeron, tomando en cuenta tres aspectos: suficiencia (que el número de ítems sean los necesarios para evaluar el concepto esperado), pertinencia (que los ítems sean precisos y acordes al tema de interés) y claridad en la redacción (que el uso del lenguaje y la terminología sean apropiados al tipo de población a quien está dirigido el instrumento).[10]

Entonces, la diferencia entre un juez y un experto es que este último si está en contacto directo con el tema de investigación y tiene suficiente experiencia en el área, aunque no necesariamente en la validación de instrumentos. Por ejemplo, si quisiéramos evaluar el conocimiento que tienen los médicos familiares en el diagnóstico de patologías benignas de la mama, los expertos para validar el instrumento serían médicos de primer contacto, especialistas en Ginecología o cualquier otro personal del área de la salud que esté directamente involucrado y en constante contacto con este tipo de pacientes. Dicho lo anterior, es importante mencionar que un juez puede ser al mismo tiempo un experto; no obstante, el hecho de que un juez no sea experto no es una condición para descartarlo del proceso de validación. [10]

Antes de que los jueces comiencen a trabajar, es imprescindible que tengan claro el fundamento teórico de la investigación y los objetivos del instrumento, es decir, el constructo. Para ello, es deseable que el investigador responsable les haga llegar por escrito la información del instrumento, así como la forma en que tendrán que realizar las observaciones, a través de diversos formatos: cartas, instructivos, listado de tareas, etcétera. Asimismo, los jueces realizarán las modificaciones y realimentarán al investigador responsable, igualmente, de forma escrita, considerando los tres elementos mencionados previamente (suficiencia, pertinencia y redacción). A esta actividad se le denomina "rondas" y debemos resaltar que no existe un mínimo o máximo de ellas; esto dependerá de la complejidad del instrumento, de la cantidad de ítems y del nivel de experiencia de todos los involucrados. Después de estas revisiones, lo más común es que los ítems y los indicadores se reduzcan, ya que los jueces hacen diferentes aportaciones al instrumento, las cuales van desde la modificación hasta la eliminación de los ítems, así como la jerarquización y eliminación de indicadores. Esto es conocido como la técnica Delphi, en la que los jueces hacen una evaluación ciega e independiente de los tallos y los reactivos, de acuerdo con la mirada teórica del instrumento, así como las respuestas a los reactivos que consideren adecuados.[13]

Otra de las tareas que se pueden llevar a cabo con ayuda de los jueces es que se evalúe que el número de preguntas por indicador sea equitativo; por ejemplo, si un instrumento es de 100 ítems y se tienen 5 indicadores, lo deseable es que se incluyan 20 ítems por indicador a evaluar, ya que de esta forma evitamos que las preguntas estén más dirigidas a un tópico que a otro. Lo anterior no es una regla para todos los instrumentos, pero es un aspecto que se puede considerar, tomando en cuenta los objetivos del instrumento en cuestión. De igual forma, algunos instrumentos también requerirán cierta homogeneidad en las respuestas de los ítems; un ejemplo son aquellos con opciones de respuesta "falso" o "verdadero", en las que lo deseable es que se construyan la mitad para ser

respondidas como falsas y la otra mitad como verdaderas.

Es así que, después de cada ronda, surge una nueva versión del instrumento y cada juez debe ser informado sobre la coincidencia de sus observaciones con otros jueces; de no ser así, se le debe solicitar que explique las razones de su opinión. Estos argumentos a favor y en contra del contenido sirven al investigador responsable para tomar decisiones sobre qué incluir o no, ya que él tiene la última palabra sobre el contenido de este. Las rondas de revisión terminan en el momento que el nivel de consenso entre los jueces es el esperado (mínimo cuatro de los cinco jueces). En este momento, se puede aplicar un test para valorar la fiabilidad intrajuez o fiabilidad interjueces, que tiene como objetivo determinar el porcentaje de acuerdo entre ellos, es decir, en qué medida coincidieron en la clasificación con relación al total de elementos examinados. A esto se le llama índice de concordancia entre evaluadores, siendo la fórmula más utilizada el índice Kappa:[14]

$$k = \frac{p_o - p_e}{1 - p_e}$$

Donde:
Po= proporción de acuerdo observado (suma de los acuerdos conseguidos en cada categoría dividida por el número de registros)
Pe= proporción de acuerdo esperado al azar (suma de la probabilidad de acuerdo por azar de cada categoría).

El resultado oscila entre 0 y 1(0.1, 0.2, 0.3, etcétera), del tal forma en que si se acerca más al uno existirá mayor concordancia.

Una vez equilibrado el instrumento, con todos los reactivos y con el índice de concordancia calculado, se procederá a la siguiente etapa.

3. Prueba premuestreo

Para este momento, el instrumento ya debe contar con la validación de contenido (versión preliminar del instrumento dada por los jueces). No obstante, es necesario saber las propiedades que tiene para medir aquello que se pretende. Entonces, requeriremos una población para aplicarlo y, posteriormente, realizar pruebas estadísticas apropiadas. A lo anterior se le conoce como la prueba premuestreo o piloto.

Lo primero será seleccionar un grupo de personas lo más parecido a la población a la que está dirigido nuestro instrumento. Por ejemplo, si éste tiene como objetivo determinar la presencia de hábitos alimenticios desfavorables en adultos mayores con diabetes, lo ideal será realizar una prueba premuestreo en pacientes que vivan con esta enfermedad y que además, sean adultos mayores. Algo que debemos resaltar es que el tamaño del grupo seleccionado para una prueba piloto no es lo más importante, pues en investigación se señala con frecuencia que las muestras grandes reducen en forma significativa la posibilidad de error, sin embargo, para los estudios que tienen como propósito realizar la validación de un instrumento, esto no es del todo cierto, debido a que lo que se toma en cuenta no es el número de participantes sino el número de preguntas o ítems que forman parte del instrumento.[15]

Normalmente, para la validación deberíamos tener entre 5 y 10 sujetos por ítem, con un mínimo de 300, ya que de esta forma se pueden tener mayores garantías respecto a la validez del instrumento; otros señalan que tener entre 2 y 3 participantes por ítem es suficiente, siempre y cuando el número total no sea inferior a 200. No obstante, se pueden permitir muestras más pequeñas si se pretende replicar la medición usando diferentes grupos, en los que el número de sujetos sea, al menos, el doble que el número de ítems, con un total no inferior a los 100 participantes por grupo.[16]

Es importante mencionar que uno de los principales propósitos de la prueba piloto es valorar la claridad del instrumento y por ello, es requisito que el investigador responsable esté presente, para que pueda resolver las dudas directamente a quienes lo contestan, en el entendido de que aún no está completamente validado.[10]

4. Validación de constructo y de criterio

Dentro de las técnicas estadísticas multivariadas se encuentra el análisis factorial (AF), utilizada frecuentemente en el proceso de validación de instrumentos. En general, se conocen dos tipos básicos de análisis factorial: el análisis factorial exploratorio (AFE) y el análisis factorial confirmatorio (AFC).

El primer tipo: AFE, tiene como objetivo tratar de establecer una estructura subyacente entre las variables del análisis, a partir de estructuras de correlación entre ellas; es decir, se agrupa a los ítems (más conocidos como factores) que estén altamente correlacionados entre sí y se les asigna un concepto.

Antes de realizar un AFE, se debe hacer una evaluación del supuesto de

correlación entre las variables, con el fin de establecer si se justifica o no su aplicación. Algunas de las estrategias más utilizadas para evaluar este supuesto es hacer una inspección de la matriz de correlaciones. Si las variables tienen valores de correlación bajos entre sí en forma general (valores menores a 0.30), es necesario cuestionar si tiene sentido este análisis. Otra alternativa para evaluar estas correlaciones es por medio de la prueba de esfericidad de Bartlett, que tiene como hipótesis nula que no existe correlación entre las variables; al rechazar esta hipótesis, se demuestra que en realidad sí existe algún grado de correlación estadísticamente significativa. Un tercer método implica evaluar la fuerza de la relación entre dos variables o ítem, utilizando el índice Kaiser Meyer Olkin (KMO), el cual toma valores entre 0 y 1; valores menores de 0,5 se consideran inaceptables; de 0.5 a 0.59, pobres; de 0.6 a 0.79, regulares, y de 0.8 a 1, aceptables.[17]

La interpretación de los resultados es uno de los aspectos más importantes del AFE, ya que depende en gran parte de la experiencia. Una de las formas es a través del método de rotación de factores que, como su nombre lo indica, significa girar los ejes factoriales a distintos grados, pero manteniendo fijo el origen, redistribuyendo la varianza de las variables originales en los factores, con el fin de lograr una mejor interpretación de los resultados. En la actualidad se utilizan dos tipos de rotaciones en AFE, que son seleccionadas por el investigador, según el conocimiento que tenga del problema. Estas rotaciones son las ortogonales y las oblicuas, de las cuales las más conocidas son Varimax, Cuartimax y Equamax (ortogonales), así como Oblimin y Promax (oblicuas). [17]

Por otra parte, el AFC corrobora que el conjunto de factores previamente organizados teóricamente (por conceptos) se ajustan. Aquí el investigador desempeña un papel muy importante, pues, a mayor conocimiento del problema, tiene mayor capacidad para formular y probar hipótesis mucho más concretas y específicas.[17]

Los dos análisis no son excluyentes, pero, dependiendo de los objetivos del instrumento, se debe decidir cuál es el más adecuado; sin embargo, en algunos casos se prefiere realizar ambos.

Una de las recomendaciones al momento de agrupar los conceptos o dimensiones es procurar un balance en la cantidad de ítems que tiene cada uno, aunque no es un requisito indispensable. Así, es posible que algunos de los ítems salgan o se agrupen en una dimensión diferente a la que inicialmente correspondían. [10]

a. Realizar la validación de criterio

Un criterio no es más que la segunda forma de evaluar el concepto que pretendemos medir. Por ello, es necesario retomar el punto 1, referente a la fundamentación teórica y empírica del instrumento.

La validez de criterio significa que los resultados obtenidos con el instrumento elaborado son similares a los que se obtienen de otros instrumentos aplicados a la misma población. Para ello, primero necesitamos saber si el concepto que se quiere medir con el instrumento ya está claramente definido en la literatura.

Si quisiéramos medir la gravedad de los síntomas del tracto urinario inferior asociados a la hiperplasia benigna de próstata (HBP), sabemos que ya existe un instrumento validado, confiable y ampliamente utilizado, que es el IPSS o puntaje internacional de síntomas prostáticos. Por este motivo, a dicho instrumento se le consideraría el estándar de oro.[18]

¿Y por qué a un instrumento se le denomina así? Porque normalmente a los autores que los elaboran les lleva mucho tiempo, muchos recursos e incluso, puede implicar que realizaron pruebas invasivas en los pacientes para poder construirlos de forma adecuada. Entonces, debemos comparar el instrumento que hemos realizado con el estándar de oro; para ello, calculamos la concordancia o correlación entre ambas escalas, la cual debe ser mayor a 0.8. Esta se puede obtener a través de la prueba estadística de concordancia Kappa de Cohen o la prueba correlación de Spearman, sobre todo si vamos a analizar las categorías del instrumento. Si, por otro lado, lo que se quiere comparar son los números obtenidos de las variables del instrumento, se utilizará el coeficiente de correlación r de Pearson.

Otra opción es que ya exista un instrumento sobre el concepto que queremos estudiar, pero no es precisamente el estándar de oro. Esto es muy frecuente en investigaciones de tipo social, educativo y psicológico. Por ejemplo, si construimos un instrumento para evaluar la autoestima, actualmente ya existen varias escalas, como la de autoestima de Rosenberg[19] y la de Coopersmith.[20] De acuerdo con lo anterior, lo que procede es evaluar a una misma población con los dos instrumentos: el que ya existe y el que nosotros construimos. Con los resultados obtenidos, también se establecerá la concordancia entre instrumentos, estableciendo una especie de consenso para la medición de dicho concepto.

No obstante, también podemos tener una tercera posibilidad: que el concepto no esté previamente definido en la literatura. Por ello, no es posible realizar una validación de criterio, pues la línea de investigación es totalmente nueva y los resultados de la aplicación del instrumento nos dirán si efectivamente, es útil para evaluar el concepto propuesto.

KR-20= $k / k - 1 [1- \Sigma piqi / \sigma T^2]$	Alfa de Cronbach= $k / k - 1 [1- \Sigma\sigma i^2 / \sigma T^2]$
k: número de ítems	*k*: número de ítems
pi: % de afirmativo del ítem	Σi: varianza del ítem
qi: complemento de p	σT: varianza total de la escala
σT: varianza total de la escala	

5. Cálculo de la confiabilidad (consistencia interna) del instrumento

Este paso consiste en calcular la consistencia interna, que se refiere al grado en que los ítems o reactivos que son parte de una escala o instrumento se correlacionan entre ellos, de tal forma que miden el mismo constructo. La confiabilidad es una medida de homogeneidad y lo esperado es que los ítems tengan una alta correlación; además, las preguntas de cada indicador también deben ser similares entre sí.[21]

a. Definir el tipo de escala utilizada

Primeramente, es importante saber qué tipo de escala se utilizó. Es así que, para aquellos instrumentos con patrón de respuesta dicotómico (por ejemplo, verdadero o falso), utilizaríamos la fórmula 20 de Kuder-Richardson (cuando los ítems tienen diferentes índices de dificultad) o 21 (cuando los índices de dificultad son iguales). Para aquellos con escala de respuesta politómica, utilizaríamos la prueba de alfa de Cronbach. Estos son dos de los métodos más comunes para calcular la consistencia interna y son equivalentes desde el punto de vista matemático.[15]

Ya teniendo los resultados de la prueba premuestreo, lo que sigue es la aplicación de las pruebas correspondientes:

Para llevar a cabo estas pruebas psicométricas, se pueden utilizar los paquetes estadísticos SPSS y STATA.

Para que la consistencia interna se considere aceptable o alta para un instrumento, debe encontrarse entre 0.70 y 0.90. Ahora explicaremos que significa este coeficiente. Por ejemplo, si un instrumento tiene una consistencia interna de 0.8, quiere decir que el 80 de la variabilidad es cierta y que 20% restante puede ser producto de un error de medición, no del instrumento en sí. Finalmente, cualquier valor inferior a 0.7 indica que existe una correlación baja entre los ítems del instrumento y por otro lado, si el coeficiente está arriba de 0.9, existe riesgo de redundancia o duplicación de ítems, debido a lo cual deberá revisarse el instrumento para que aquellos que están duplicados se eliminen.[15]

Es común que instrumentos que tienen más de 20 ítems, tengan una consistencia interna mayor a 90, ya que cuando hacemos la sustitución en la fórmula correspondiente, se debe poner el número de ítems. Por ello, es recomendable calcular también la consistencia interna por grupos de ítems, con lo cual se evitará una sobreestimación del instrumento.[10]

Uno de los métodos que se puede emplear es el de semipartición (división del instrumento en dos mitades), de modo que tengan el mismo número de ítems cada mitad y que puedan ser consideradas paralelas. Posteriormente, se calcula la puntuación total en cada una de estas partes. Un ejemplo de cómo separar las mitades podría ser en un instrumento de 200 ítems: una mitad serían los 100 primeros y la segunda los 100 restantes, o un grupo de ítems serían los pares y el otro, los ítems impares. Así, para calcular la consistencia entre grupos de ítems se puede utilizar la fórmula de corrección de Spearman-Brown.[22]

$$r_{xx'} = \frac{2r_{AB}}{1 + r_{AB}}$$

Esta fórmula expresa la relación entre la longitud y la fiabilidad del instrumento, bajo el supuesto de que ambas partes del mismo son paralelas. Aquí, r es el coeficiente de confiabilidad para la mitad de la prueba, rxx' es la confiabilidad para la prueba total; asimismo, tenemos la puntuación en la forma A y en la forma B para cada sujeto.[22]

Poniendo un ejemplo, si la correlación de las puntuaciones totales de los ítems impares con las puntuaciones

totales de los ítems pares es 0.85, la confiabilidad estimada de toda la prueba sería:

$$r_{xx'} = \frac{2r_{AB}}{1 + r_{AB}}$$

b. Calcular el índice de correlación

Una vez obtenida la consistencia interna de todos los ítems (tanto general como por grupos de ítems), se ordenan de acuerdo con su índice de correlación, de los que tienen mayor a los que tienen menor correlación. Siguiendo el ejemplo anterior, si tenemos un grupo de 200 ítems, se eliminarán aquellos con menor magnitud de correlación (consistencia menor a 0.8). Una posible explicación de por qué los ítems pueden tener una baja correlación es porque son ambiguos. Si aún se siguen teniendo ítems con baja correlación, no será lo más adecuado seguirlos eliminando, sino implementar un método para aumentar el valor alfa de Cronbach. Para ello, es necesario ordenar los ítems, pero no de acuerdo con su índice de correlación, sino de acuerdo con su varianza.

En la primera fila se coloca el ítem que tiene el mayor grado de variabilidad o mayor magnitud de la varianza y hacia abajo, en orden, los que tengan el menor grado de variabilidad. Ya ordenados, se seleccionan los primeros ítems con menor grado de variabilidad y se modifica la redacción con la finalidad de que la forma en que contestan los sujetos sea más dispersa; se vuelve a calcular el valor alfa de Cronbach. De no modificarse, pasaríamos con los ítems que siguen, en orden ascendente. Al ir realizando las modificaciones a los ítems para evitar que todos los sujetos contesten lo mismo, el alfa de Cronbach se verá beneficiado.[10] Este mismo método se puede emplear con la prueba de Kuder-Richardson, siguiendo los pasos comentados previamente.

Conclusión

Este estudio enfatiza la importancia crítica de validar adecuadamente los instrumentos utilizados en la recolección de datos clínicos para asegurar la precisión y fiabilidad en el diagnóstico y tratamiento médico. La aplicación rigurosa de métodos como la clinimetría, la validación por jueces expertos y el análisis factorial, tanto exploratorio como confirmatorio, permite confirmar que estos instrumentos cumplen con los estándares necesarios para medir eficazmente variables de interés.

Referencias

1. Villasís KM, Márquez GH, Zurita CJ, Miranda NM, Escamilla NA. El protocolo de investigación VII. Validez y confiabilidad de las mediciones. Rev Alerg Mex. 2018;65(4):414-21.
2. Iglesias GA, Quintana G. Análisis histórico de la clinimetría y de la autoclinimetría. Estado del arte. Rev Colomb Reumatol. 2013;20(1):1–8.
3. Organización Mundial de la Salud [Internet]. [Citado 2023 Jul 5]. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight
4. Padrós BF, Montoya PK, Bravo CM, Martínez MM. Propiedades psicométricas del Inventario de Ansiedad de Beck (BAI, Beck Anxiety Inventory) en población general de México. 2020; 26:181–187.
5. Dos Santos ERP, Coelho JCF, Ribeiro I, Sampaio F. Translation, cultural adaptation and evaluation of the psychometric properties of the Hamilton Anxiety Scale among a sample of Portuguese adult patients with mental health disorders. BMC Psychiatry. 2023;23(1):520.
6. Hernández SR, Fernández CC, Baptista LM. Metodología de la investigación. 6a ed. México: McGraw-Hill Interamericana; 2014.
7. Soriano AM. Diseño y validación de instrumentos de medición. Diálogos. 2014;14,19-40.
8. Kerlinger F. Investigación del Comportamiento. 4a ed. México: McGraw-Hill; 2002.
9. Bravo PT, Valenzuela GS. Desarrollo de instrumentos de evaluación: cuestionarios. Chile: Centro de Medición MIDE UC; 2019.
10. Supo J. Cómo validar un instrumento. La guía para validar un instrumento en 10 pasos. Perú: 2013 [Internet]. [Citado 2023 Ago 22]. Disponible en: https://dspace.uniandes.edu.ec/handle/123456789/16000
11. Urzúa MA. Calidad de vida relacionada con la salud: elementos conceptuales. Rev Med Chile. 2010;138(3):358-365.
12. Matas A. Diseño del formato de escalas tipo Likert: un estado de la cuestión. REDIE. 2018;20(1):38-47
13. Cobos AH. Cómo construir un instrumento para evaluar la lectura crítica de investigación de informes médicos. Inv Ed Med. 2021;10(39):96-105.
14. Dubé JE. Evaluación del acuerdo interjueces en investigación clínica. Breve introducción a la confiabilidad interjueces. Revista Argentina de Clínica Psicológica. 2008; XVII(1):75-80.
15. Campo AA, Oviedo HC. Propiedades psicométricas de una escala: la consistencia interna. Rev Salud Pública. 2008;10(8):831-839.
16. Roco VA, Hernández OM, Silva GO. ¿Cuál es el tamaño muestral adecuado para validar un cuestionario? Nutr. Hosp. 2021;38(4):877-878.
17. Méndez MC, Rondón SM. Introducción al análisis factorial exploratorio. Rev Colomb Psiquiatr. 2012;41(1):197-207.
18. Preciado ED, Kaplan SA, Iturriaga GE, Ramón TE, Mayorga GE, Auza BA, et al. Comparación del Índice Internacional de Síntomas Prostáticos versus Escala Visual Análoga Gea para la evaluación de los síntomas de la vía urinaria inferior. Rev Mex Urol. 2017;77(5):372-382.
19. Martínez RG, Alfaro UA. Validación de la escala de autoestima de Rosenberg en estudiantes paceños. Fides Et Ratio. 2019;17(17):83-100.
20. Díaz RA, Pérez MG, Puentes ML, Castillo MM. Fiabilidad y validez de constructo del Inventario de Autoestima de Coopersmith en estudiantes de medicina. Revista de Ciencias Médicas de Pinar del Río. 2022; 26 (3):e5371.
21. Rodríguez RJ, Reguant AM. Calcular la fiabilidad de un cuestionario o escala mediante el SPSS: el coeficiente alfa de Cronbach. REIRE Revista d'Innovació i Recerca en Educació. 2020;13(2):1–13.
22. Warrens MJ. Transforming intraclass correlation coefficients with the Spearman-Brown formula. J Clin Epidemiol. 2017;85:14-16.