Validez en la evaluación de escalas

Julio Alejandro Lamprea M.[1]
Carlos Gómez-Restrepo[2]

Resumen

Las escalas como instrumento de medición adquieren particular importancia cuando la variable o entidad que se pretende medir es subjetiva, es decir, que tanto su definición como sus componentes tienen un alto nivel de variabilidad. Dentro de las características que deben estar presentes en cualquier instrumento de medición se incluyen su reproducibilidad, utilidad y validez. La validez permite hacer inferencias y evaluar hipótesis acerca de las personas sobre las cuales estamos empleando las escalas como instrumento de medición. El propósito de este artículo es presentar la importancia y la utilidad de la validez en el proceso de evaluación de escalas. Se describen las diferentes clases de validez y su forma de aplicación en un contexto clínico, para determinar la presencia de este atributo como parte del proceso de validación de escalas.

Palabras clave: estudios de validación, medición.

Title: Validity in Scale-testing.

Abstract

Scales are of particular importance when the variable or the entity that we pretend to measure is subjective, that is, there is a high degree of variability among its definition and its components. Among the qualities that a measurement instrument must have are its utility, reliability, and validity. Validity allows us to test hypothesis and make inferences about the people being administered with the scale as a measurement instrument. The purpose of this article is to show the importance and the utility of validity in the process of scale-testing. The different classes of validity and their application in a clinical context in order to determine their presence are discussed.

Keywords: Validation studies, measures.

[1] Médico. Asistente de investigación del Departamento de Epidemiología Clínica y Bioestadística, Facultad de Medicina, Pontificia Universidad Javeriana, Bogotá, Colombia.
[2] Médico psiquiatra, psicoanalista, MSc. en Epidemiología Clínica, profesor asociado del Departamento de Epidemiología Clínica y Bioestadística y del Departamento de Psiquiatría y Salud Mental, coordinador del Programa Psiquiatría de Enlace de la Facultad de Medicina, Pontificia Universidad Javeriana, Bogotá, Colombia.

Introducción

Una escala puede definirse como la "colección de ítems que pretenden revelar diferentes niveles de determinadas características (variables) no observables directamente. Estas escalas se desarrollan cuando se quiere medir fenómenos no directamente observables" (1). Cantidades físicas como medición de glucemia, concentración de hormona tiroidea, enzimas cardiacas, talla, peso o índice de masa corporal son observables directamente o con los instrumentos adecuados (2).

Esta medición objetiva de ciertas variables deja poco espacio a la duda de si en realidad lo que intentamos medir corresponde a dicha variable. Cuando, por ejemplo, se le toma la presión arterial a una persona, independientemente de si el instrumento de medición está bien calibrado o no, la pregunta de si en realidad lo que estamos midiendo es la presión arterial de este paciente y no otra variable, como temperatura o pulso, tiene una respuesta obvia: lo que estamos midiendo es la presión arterial.

Desafortunadamente en medicina —y especialmente en psiquiatría— es frecuente que no encontremos una medida o un instrumento de medición único que nos dé la certeza de que el atributo o la condición que pretendemos medir corresponde a la realidad.

El problema radica en que las variables con que nos encontramos frecuentemente, así como su medición, "dependen tanto de sus definiciones, que varían de persona a persona, y de su forma de medición, como de la relación entre la observación y lo que ésta refleja" (2). Por ejemplo, para evaluar la prevalencia del alcoholismo en una población o para determinar el riesgo individual de tener dependencia o abuso de alcohol se han elaborado diferentes escalas, cada una con una serie de preguntas que buscan determinar la presencia de alcoholismo.

Debido a que no existe uniformidad de criterios en cuanto a la definición ni a la forma de medir esta entidad, así como tampoco se tiene certeza de que ciertas conductas observadas en la población en que se quiere medir o determinar abuso o dependencia de alcohol (por ejemplo consumo de alcohol matutino en la escala CAGE de tamizaje para alcoholismo) sean predictivas de la presencia de tal abuso o dependencia, cada una de estas escalas ha tenido que ser evaluada para determinar si en realidad los ítems que en ellas se incluyen son predictores de la presencia de abuso o dependencia de alcohol, es decir, si estas escalas son instrumentos de medición válidos.

Aunque el objetivo de este artículo es examinar la validez en el proceso de evaluación de escalas, es necesario tener en cuenta que en cualquier instrumento de medición existen otras características además de la validez que deben ser evaluadas. Estas características incluyen: confiabilidad, adecuada amplitud de rango, sensibilidad al cambio y utilidad del instrumento

de medición (3). Por esta razón, la validez debe ser entendida como uno de los atributos que deben ser evaluados en diferentes instrumentos de medición.

Definición

La validez o exactitud de una escala se refiere al "grado de confianza que podemos tener de que la medición corresponde a la realidad del fenómeno que se está midiendo" (1). Si, por ejemplo, tenemos los resultados de una nueva escala que busca determinar la existencia de depresión en una persona, y se nos informa que los pacientes evaluados obtienen los mismos puntajes en diferentes ocasiones y que diferentes entrevistadores obtienen resultados similares al evaluar al mismo paciente, podemos admitir que la escala es confiable o reproducible; sin embargo, con estos datos no podemos determinar si la escala mide la presencia de depresión y no otra condición, como estrés, demencia o ansiedad.

Aunque necesaria e indispensable en el proceso de validación de una escala, la reproducibilidad no es, por sí sola, suficiente para determinar la validez de una escala. La validez determina qué conclusiones pueden derivarse de las personas que obtienen diferentes resultados en la escala (4). Por lo tanto, si en nuestra hipotética escala una persona obtiene un puntaje alto, ¿qué tanta confianza podemos tener de que esta persona, al haber obtenido este puntaje, tenga de hecho depresión y no otra condición clínica? Otra forma de definirla sería: la validez es el grado en el cual la evidencia y la teoría soportan las interpretaciones obtenidas de los resultados de las pruebas (escalas, inventarios, etc.), siendo la validez un concepto unificado, de tal forma que, más que tipos discretos de validez, existen tipos de evidencia de validez (5,6).

Al aceptar que los procesos de validación de escalas están encaminados a las inferencias que se pueden hacer acerca de las características de las personas que han obtenido diferentes puntajes en estas escalas, "la validez de una escala resulta ser un proceso de evaluación de hipótesis" (2).

Retomando nuestro ejemplo de la nueva escala de depresión, al determinar su validez estaríamos evaluando la hipótesis según la cual la escala nos permite concluir que alguien que obtiene un puntaje alto presenta, de hecho, depresión; además, permite diferenciar a esta persona de otras con diferentes entidades y de personas que obtengan puntajes bajos en esta escala.

De acuerdo con esto, al querer determinar la validez de una escala, las preguntas necesarias que se deben hacer son: ¿la hipótesis planteada en este proceso de validación guarda relación con lo que la escala busca medir? y ¿los resultados de esta escala nos permiten hacer las inferencias que queremos? (2).

Tradicionalmente la validez de una escala se ha dividido en: (a) validez de apariencia, (b) validez de contenido, (c) validez de criterio y (d) validez de constructo. Estos

componentes, los cuales describiremos a continuación, no deben ser entendidos como entidades totalmente diferentes e independientes unas de otras, sino como criterios que intentan establecer el grado de confianza que podemos tener de las inferencias que realicemos acerca de las personas que obtengan puntajes en escalas (2).

a. Validez de apariencia

La validez de apariencia tiene como objetivo responder a la siguiente pregunta: ¿la escala parece medir lo que debe medir? La validez de apariencia "no supone un concepto estadístico, sino que depende de los juicios que los expertos hagan sobre la pertinencia de los ítems de la escala" (7).

Para evaluar la validez de apariencia se conforma un grupo de jueces, por lo general expertos que determinan si en su concepto el instrumento en apariencia mide las cualidades deseadas, y otro de personas que van a ser evaluadas por la escala.

La importancia de esta forma de validez radica en la aplicabilidad y en la aceptabilidad desde el punto de vista del que responde a la escala (4). Por lo tanto, una persona puede estar más dispuesta a responder un cuestionario de una escala que quiera determinar ansiedad, si los ítems aparentemente están abordando de alguna forma la presencia de esta entidad.

Por el contrario, si en la escala aparecen otros ítems que puedan ser juzgados por parte del que responde como irrelevantes, entonces esto hará que el cuestionario pierda interés para la persona evaluada y que posiblemente esté midiendo otra entidad o constructo. Un ítem que dice "La mayoría del tiempo me siento triste" en apariencia mide depresión; no obstante, otro que dice "Siento como si mi pensamiento fuera audible por las demás personas" puede corresponder más a un trastorno esquizofrénico que a uno depresivo. En apariencia, mediría otra entidad.

b. Validez de contenido

Cuando se desarrolla o se quiere evaluar una escala, se debe pretender que los ítems en esta escala cubran adecuadamente todos los dominios de la entidad que se quiere medir. Un dominio es un grupo de características que se encuentran comúnmente presentes en la entidad, y los ítems son herramientas de exploración que nos permiten evaluar la presencia de estos dominios.

En la escala de Glasgow, por ejemplo, que se utiliza en la evaluación inicial de pacientes con trauma craneoencefálico y cuyo objetivo fundamental es establecer el pronóstico de estos pacientes, se establecieron tres dominios principales (respuesta motora, respuesta verbal y apertura ocular). Cabe aclarar que cada uno de estos dominios es explorado por ítems que establecen el puntaje total de la escala.

Si, para ahorrar tiempo o procurando hacer más simple la anterior escala, removiéramos uno de sus dominios, como la evaluación

de la respuesta motora, la inferencia acerca del pronóstico de esta persona basado únicamente en su puntaje en la respuesta verbal y apertura ocular dejaría de ser válida, ya que estaríamos eliminando de la escala una evaluación necesaria para realizar esta inferencia. Es decir, la escala dejaría de tener validez de contenido o, por lo menos, este tipo de validez se vería comprometida.

En el proceso de construcción de una escala se debe procurar que los dominios incluidos y los ítems para explorarlos representen adecuadamente a la entidad que se va a medir; de esta forma se logra que las inferencias surgidas a partir de puntajes en la escala sean válidas dentro de un rango amplio de circunstancias.

En el caso de la calidad de vida existen escalas genéricas, como el SF-36, que poseen una amplia cantidad de subescalas o dominios: funcionamiento físico, desempeño físico, dolor corporal, desempeño emocional, salud mental, vitalidad, salud general, funcionamiento social y, en algunos, cambio de la salud en el tiempo. Esto nos hace pensar en la necesidad de medir todas estas subescalas con el fin de tener completo el constructo de calidad de vida.

Cuando la entidad que se quiere medir es muy heterogénea, surge el problema de que inevitablemente la reproducibilidad de los resultados de la escala se verá comprometida. Por ejemplo, en el lupus eritematoso sistémico, no todos los pacientes que exhiben manifestaciones hematológicas presentan afectación renal o cutánea; es decir, la entidad presenta poca consistencia interna (una medida de reproducibilidad), lo que podría solucionarse incluyendo en la escala únicamente dominios que aparezcan en conjunto con alta frecuencia, mejorando de este modo la confiabilidad de la escala.

Si esto se hiciera, sin embargo, estaríamos comprometiendo la validez de contenido de la escala, puesto que los dominios incluidos que mostrarían mayor consistencia interna no abarcarían el espectro real de la entidad, afectando de esta manera las inferencias que hagamos de la escala. Cuando se presenta esta disparidad entre confiabilidad de la escala y su validez de contenido, resulta más adecuado sacrificar en algo la primera, con objeto de mantener la validez de contenido; todo esto, si aceptamos que el principal objetivo es inferencial, lo cual depende más de la validez de la escala (2). No obstante, una escala válida pero no confiable de poco serviría, puesto que su no reproducibilidad la haría poco útil en el ámbito clínico. De esta forma, es fundamental procurar obtener ambas.

c. Validez de criterio

Para establecer si los puntajes obtenidos a partir de una escala son válidos, ésta debe compararse con una forma de medición previamente existente (patrón de oro) que a su vez haya mostrado ser el mejor instrumento disponible para la medición de la entidad. Cuando realizamos esta comparación y aceptamos que existe una adecuada correlación entre estos dos instrumentos de medición, estamos asegurando que

la escala tiene validez de criterio.

En un estudio clínico reciente (8) se pretendió validar una escala que incluía medidas físicas, como índice de masa corporal y circunferencia abdominal junto con datos simples de laboratorio —recuento de plaquetas y concentración de glucosa en sangre— para establecer el diagnóstico de enfermedad hepática grasa de origen no alcohólico con fibrosis avanzada (NAFLD, por su sigla en inglés). El criterio o patrón de oro que se usó para comparar los puntajes obtenidos de la escala fue la biopsia hepática.

La razón para querer validar esta escala como instrumento de medición, respecto a otro que ya había demostrado ser preciso, es que si se demostrara que esta escala se correlaciona adecuadamente con el patrón de oro, resultaría entonces más cómodo y menos riesgoso para el paciente, así como más económico, y no estaría sujeta a errores de muestra, a diferencia de la biopsia. Estos atributos (menores costos, menor riesgo para el paciente, procedimiento no invasivo, mayor simplicidad en la aplicación) (4) son los que se deben buscar cuando se quiere remplazar un instrumento de medición por otro ya existente.

En el anterior estudio, debido a que la escala y el criterio (biopsia) se aplicaron al mismo tiempo, a la validez evaluada se le llama *validez concurrente*. Dado que la variable anterior es dicotómica (presencia de fibrosis hepática o no), los datos se pueden agrupar en una tabla de 2 x 2, y se pueden analizar los resultados usando índices como sensibilidad o especificidad, o valor predictivo positivo o negativo. Si las variables fueran continuas, sería necesario usar el coeficiente de correlación de Pearson (7).

En cualquier caso, lo que se debe buscar es que exista una adecuada correlación entre los dos instrumentos de medición. Si la razón para aplicar el nuevo instrumento es mayor beneficio en los atributos anteriormente discutidos, con una validez similar al criterio, se buscan índices de correlación mayores a 0,8 (es decir, se necesita que exista una alta correlación). Si, por el contrario, se quisiera crear una nueva escala más válida que el anterior instrumento de medición, se debería pretender obtener correlaciones entre 0,3 y 0,7; en primer lugar, que la correlación no sea tan alta, asegurando que los dos instrumentos no estén evaluando el mismo atributo, con lo que se dejaría poco espacio para "que el nuevo instrumento sea mejor y diferente del original"; en segundo lugar, que no sea tan baja como para pensar que los instrumentos no están relacionados, "queriendo decir que la escala está midiendo algo completamente diferente del criterio" (4). Un tipo de validez concurrente en psiquiatría sería evaluar la entrevista diagnóstica estructurada (por ejemplo, CIDI 2.1) para depresión, que actuaría como patrón de oro, frente a los resultados del Hamilton, Beck o alguna otra escala para depresión, que se aplica al mismo individuo. Se esperaría que hubiera una alta correlación entre los resultados que

arroje el CIDI y aquel del Hamilton para depresión.

En el caso de que el criterio o patrón de oro no sea evaluado al mismo tiempo que la nueva escala, sino en algún punto en el futuro, se hablaría de validez predictiva. La validez del nuevo instrumento de medición radicaría en qué tan bien predice el puntaje del criterio. Correlacionar puntajes en la escala de potencial suicida con intentos de suicidio en el futuro (criterio) (4), o correlacionar una escala que indique adherencia al tratamiento con el posterior desarrollo de esta adherencia o no (criterio) (1), o una escala que detecte demencia de Alzheimer incipiente con el posterior diagnóstico de demencia de este tipo, son ejemplos de esta clase de validez.

Es necesario aclarar que mientras no se tengan los resultados de la correlación en algún punto en el futuro, ninguna decisión se puede hacer basándose en la nueva escala, del mismo modo que no se pueden usar atributos del nuevo instrumento como parte del criterio.

Si, por ejemplo, para validar el uso de una ecografía abdominal para diagnosticar sangrado abdominal, usáramos el juicio clínico como patrón de oro o criterio, el médico podría saber los resultados de la ecografía, y basar su juicio clínico en los resultados de ésta. Esto significaría que el diagnóstico o criterio estaría basado, en parte, en las predicciones del nuevo instrumento de medición, con lo que se establecería una "alta correlación artificial entre los dos", lo que se conoce como *contaminación del criterio* (2). Por este motivo, la medición de ambas debería idealmente ser ciega, con el fin de que el conocimiento de una no influya sobre lo que detectamos en la segunda.

d. Validez de constructo

En el estudio de la esquizofrenia con el efecto que tienen los neurolépticos de disminuir la neurotransmisión de la dopamina, que a su vez produce una disminución de los delirios y alucinaciones experimentados por los pacientes, se instauró la teoría dopaminérgica de la esquizofrenia, que pretendía explicar las diferentes manifestaciones de esta entidad con fundamento en niveles anormales de este neurotransmisor. Sin embargo, pronto se encontraron las múltiples limitaciones de esta teoría (9).

Las variadas acciones de la dopamina en el sistema nervioso central (SNC), las concentraciones normales de este neurotransmisor en personas con esquizofrenia y la falta de respuesta de ciertos síntomas al tratamiento con neurolépticos llevaron a plantear nuevas hipótesis que pudieran explicar de una forma más completa las múltiples manifestaciones de esta entidad. Es decir, lo anterior ha conducido al desarrollo de nuevas "teorías" que, a su vez, explican la asociación entre varios de los síntomas de esta entidad, o al desarrollo de constructos teóricos que cumplieran este fin.

Como se discutió, muchos de los eventos que queremos medir en medicina (ansiedad, depresión, esquizofrenia) no son observables

directamente, ni existe una prueba única que nos indique precisamente la presencia de esta entidad. Por lo tanto, lo mejor que podemos hacer para medir estas variables abstractas es inferir que ciertos atributos o manifestaciones de la entidad se correlacionan para explicar la variable que deseamos medir.

Las razones para querer implementar nuevas teorías o constructos que nos permitan realizar estas inferencias son, en primer lugar, que no exista un instrumento de medición que evalúe la condición (por ejemplo, si quisiéramos desarrollar un instrumento que nos permitiera reunir varios dominios presentes en personas con riesgo de cometer asesinatos en serie o masivos) y, en segundo término, que el instrumento de medición existente sea juzgado como inválido o impreciso (4). Esta última razón difiere de la validez de criterio en que el nuevo instrumento de medición no se desarrolla por razones de costo o de riesgo para el paciente, sino con el objetivo de configurar un nuevo instrumento más adecuado o preciso que el anterior.

Para establecer la validez de un constructo se debe evaluar y probar la validez de cada parte de éste, es decir, evaluar cada una de las predicciones que se pueden hacer sobre el constructo (4). Volviendo a la teoría dopaminérgica de la esquizofrenia, una de nuestras predicciones podría ser que los medicamentos antidopaminérgicos causen una remisión de los síntomas o que personas con esquizofrenia muestren niveles anormales de este neurotransmisor.

Para validar nuestro constructo hipotético, debemos evaluar cada una de las predicciones realizadas a partir del constructo, que de ser ciertas nos harían tener mayor confianza en nuestra teoría, pero que nunca nos llevarían a tener una certeza total acerca de su validez, pues si probáramos que una de nuestras predicciones resulta falsa, le restaría validez al constructo. De otra forma, al validar un constructo no sólo estamos validando un instrumento de medición, sino la teoría al mismo tiempo (2).

Por lo tanto, si son ciertas las predicciones que hacemos a partir de la teoría, estaríamos validando tanto el constructo como el instrumento de medición. Si, en cambio, la nueva escala no nos permitiera discernir entre individuos con diferentes puntajes en ésta, tanto la teoría como el instrumento podrían estar errados, o tal vez sólo uno de ellos (2).

Conclusión

Al evaluar un instrumento de medición (escalas en este caso) debe procurarse que cumpla con una serie de atributos necesarios para poder ser validado. Estos atributos incluyen: confiabilidad, sensibilidad al cambio, adecuada amplitud de rango, utilidad y validez. Esta última característica nos permite generar un grado de confianza con el cual podamos asegurar que lo que pretendemos medir corresponde a la realidad, y que las inferencias

acerca de los puntajes obtenidos por las personas evaluadas por estos instrumentos sean confiables.

La validez adquiere particular importancia cuando los eventos que buscamos medir no son identificables directamente, sino que se construyen a partir de ciertos atributos, conductas o síntomas que llamamos *dominios*, los cuales se pretende que sean predictivos o que puedan medir adecuadamente una condición determinada.

La validez puede clasificarse en validez de apariencia, validez de contenido (que pretende que los dominios incluidos en el instrumento de medición representen adecuadamente el espectro de la entidad), validez de criterio (donde se encuentra la correlación del nuevo instrumento con un criterio o patrón de oro, de manera concurrente y de forma predictiva) y validez de constructo (que evalúa la precisión de una nueva teoría que remplace un instrumento de medición existente, o que cree uno nuevo cuando no se cuente con él).

Aunque resulta más simple realizar esta división para entender el concepto de validez, es necesario no desviarse del objetivo principal de este atributo: permitirnos evaluar hipótesis a partir de los resultados obtenidos con el instrumento de medición (2).

Referencias

1. Gómez C, Sánchez R. Conceptos básicos sobre validación de escalas. Rev Colomb Psiquiatr. 1998;27:121-30.
2. Streiner DL, Norman G. Validity: Health measurment scales. A practical guide to their development and Use. Oxford: Oxford University Press, 2nd ed.; 1995.
3. Gómez C, Ospina MB. Adaptación y validación de escalas. En: Ruiz A, Gómez C, Londoño D. Investigación clínica: epidemiología clínica aplicada. Bogotá: Centro Editorial Javeriano; 2001.
4. Streiner DL. A checklist for evaluating the usefulness of rating scales. Can J Psychiatry. 1993;38:140-8.
5. Goodwin LD. The meaning of validity. J Pediatr Gastroenterol Nutr 2002;35:6-7
6. Goodwin LD. Changing conceptions of measurement validity: an update on the new standards. J Nurs Educ. 2002;41:100-6.
7. Sánchez R, Echeverry J. Validación de escalas de medición en salud. Rev Salud Pública. 2004;6:302-18.
8. Angulo P et al. The NAFLD fibrosis score: a noninvasive system that identifies liver fibrosis in patients with NAFLD. Hepatology. 2007; 45:846-54.
9. Freedman R. Schizophrenia. N Engl J Med. 2003;349:1738-49.

Recibido para evaluación: 4 de abril de 2007
Aceptado para publicación: 22 de mayo de 2007

Correspondencia
Julio Alejandro Lamprea
Departamento de Epidemiología Clínica y Bioestadística
Hospital Universitario San Ignacio
Cra. 7 N.° 40-62, piso 2. Bogotá, Colombia
jlamprea@javeriana.edu.co