Conceptos Básicos de Validación de Escalas en Salud Mental

MARÍA DE LOS ANGELES RODRÍGUEZ GAZQUEZ[1]
JOSEFINA LOPERA JARAMILLO[2]

INTRODUCCION

Desde los albores de la humanidad el hombre se ha preocupado por medir los procesos relacionados con la salud y la enfermedad, la historia cuenta como la cultura Asirio-Babilónica en el año 668 a 626 A.C manejó estadísticas cualitativas, las que podían consultarse en la biblioteca del Rey Assurbamipal. Por cientos de años las naciones del mundo han estado generando datos estadísticos con el fin de describir la situación de salud de sus poblaciones.

Esta información permite fundamentalmente analizar las condiciones sociales y económicas de los países a través de indicadores que cuantifican los fenómenos en diferentes niveles: educativo, laboral, salud y enfermedad, perfiles de morbilidad, mortalidad, factores de riesgo y son el pilar fundamental para la toma de decisiones y la formulación de políticas de los gobiernos. Sin embargo la medición de la salud ha tenido algunas dificultades, ya que los indicadores no reflejan satisfactoriamente las condiciones de salud en los grupos de mayor exposición a riesgos y además van a ser objeto de priorización en Salud Pública.

A partir del momento en que la Organización Mundial de la Salud (OMS) define la salud como el "completo bienestar físico, mental y social, y no solamente la ausencia de enfermedad", y precisa la Salud Mental como "un estado sujeto a fluctuaciones provenientes de factores biológicos y sociales en que el individuo se encuentre en condiciones de seguir una síntesis satisfactoria de sus tendencias instintivas, potencialmente antagónicas, así como de formar y sostener relaciones armónicas con los demás y participar constructivamente en los cambios que puedan introducirse en el medio ambiente físico y social"; los epidemiólogos en todo el mundo se dieron a la tarea de desarrollar técnicas que midieran estos aspectos, incluyendo el bienestar. Estas definiciones han sido objeto de numerosas y calurosas discusiones, en el sentido de poder reflejar satisfactoriamente su significado a través de instrumentos que permitan su medición, esto precisamente es la validación: evaluar si el instrumento mide lo que debe medir.[1]

[1] Especialista en Epidemiología, Subespecialista en Salud Mental.

[2] Magíster en Salud Pública, Subespecialista en Salud Mental, Jefe de la Dirección de Planeación y Autoevaluación. Instituto de Ciencias de la Salud – CES.

La medición de los aspectos que se refieren a la Salud Mental difieren ampliamente de la de aspectos orgánicos, ya que los segundos pueden tener formas "objetivas" de evaluar la salud, en cambio en los primeros estas mediciones pueden ser muy "subjetivas".

Durante la segunda guerra mundial se creó la necesidad de evaluar la salud mental en un número grande de soldados donde era imposible la evaluación individual; por esta razón se crearon tests o escalas, que fueron puntuados y después estandarizados para este fin.[2]

Otra contribución del período de guerra fue la aplicación de escalas numéricas a los índices de salud, debido a que estos reportes no eran exclusivamente cuantitativos, algunos métodos fueron adquiridos para evaluar categóricamente preguntas tales como "siento un dolor severo" en una forma práctica para hacer el análisis estadístico.

Originalmente las técnicas de medición de las escalas fueron desarrolladas por la psicología social con el fin de medir actitudes y su uso permitió sistematizar los procesos de la medición subjetiva. Lo anterior estableció las bases de medición para los eventos de salud mental en las poblaciones.[3]

Algunos autores han tratado de clasificar la medición en salud de varias formas; una de ella es la funcional, la cual distingue la medición según su propósito de diagnóstico, pronóstico o evaluación. Dentro de éstos tests están las escalas para evaluar eventos de salud mental con cualquiera de los tres fines.[4]

Otra clasificación muy frecuentemente utilizada es la específica del evento, de acuerdo al objeto de la medición del evento a evaluar: depresión, ansiedad, calidad de vida en pacientes terminales, satisfacción del usuario, etc.; cada uno de ellos tendrá una escala para su medición.

Antes de revisar los procedimientos de evaluación de la validez de las escalas de medición en salud, se necesita resolver la pregunta: ¿Qué evidencia existe que permita un juicio subjetivo de la medición en salud?

En contraposición con la tradicional forma de medir en ciencias exactas, no se puede decir que la evidencia que proporciona este tipo de medición sea débil. Para introducir las bases científicas de la medición en salud es necesario empezar con una descripción de los procedimientos psicométricos utilizados para asignar puntajes numéricos a juicios subjetivos. Los argumentos que consideran estos juicios subjetivos como válidos provienen de los estudios de la psicofísica y de las técnicas psicométricas utilizadas para medir la salud.

1. ESTIMACION NUMERICA DE LAS MEDICIONES EN SALUD: METODOS PARA DISEÑAR ESCALAS

Las mediciones de juicios "subjetivos" pueden estimarse cuantitativamente a través de instrumentos denominados "escalas".

La escala es una herramienta que permite asignar valores a ítems de objetos de acuerdo a una regla de decisión. Usualmente los valores elegidos son números y las opciones de respuesta de los ítems se definen previamente en tal forma que sean mutuamente excluyentes y exhaustivas.[5]

Para elegir la escala adecuada en investigación, debe llevarse a cabo una búsqueda completa de la literatura a través de diversas bases de datos electrónicos y posteriormente complementándola con la búsqueda manual.

Una vez se hayan localizado una o más escalas relacionadas con el fenómeno a evaluar, es importante tomar la decisión de sí es posible utilizar el instrumento elegido o es necesario diseñar un nuevo instrumento. Esta decisión debe ser analizada con cuidado para emitir juicios apropiados en la selección de los ítems de la escala y por esta razón es importante tener en cuenta una valoración crítica adecuada de la evidencia que apoye el instrumento.

Las escalas se realizan en general con el fin de:

- **Probar hipótesis**: Estas pueden ser de tipo: "unidimensional" refiriéndose a un solo constructo latente de interés o "multidimensional"más complejas en el sentido de referirse a dos o más constructos.
- **Análisis exploratorio**: Para determinar qué dominios subyacen en la escala.
- **Asignación de puntuación**: Asignación de valores o puntajes a los ítems de los dominios.

2. CONSIDERACIONES PARA LA ELABORACION DE LOS ITEMS DE UNA ESCALA

El primer paso para la elaboración de los ítems de una escala, es analizar lo que otros investigadores han realizado en el pasado y han considerado como relevante, importante o discriminativo. Además, hay que analizar si los ítems específicos que fueron utilizados para desarrollar tests individuales, provienen de tests anteriores.

Hay algunas razones del por qué los ítems son utilizados nuevamente en otras escalas, partiendo de inventarios previos: **Primero**, ahorra trabajo ante la necesidad de construir nuevas escalas. **Segundo**, estos ítems usualmente han pasado por procesos de validación que garantizan su capacidad psicométrica; y por último: hay un número limitado de formas de hacer las preguntas sobre problemas específicos, por ejemplo si nuestro interés consiste en evaluar el estado depresivo, es difícil preguntar acerca de la pérdida del sueño en una forma que no haya sido utilizada con anterioridad.

2.1. FUENTES DE INFORMACION PARA OBTENER LOS ITEMS DE UNA ESCALA

A través de los años, una gran variedad de técnicas han sido desarrolladas para elegir las fuentes de información de escalas específicas, resaltando que en cada una de ellas los pacientes y sujetos de una investigación son la fuente potencial de los ítems. Inicialmente estos procedimientos se elaboraron partiendo de investigación cualitativa y sólo en los últimos años se han utilizado abordajes cuantitativos para este fin. Siendo las técnicas de mayor relevancia según Willms y Johnson: [6]

2.1.1 Grupos focales: Son grupos de informantes constituidos por seis a doce personas, guiadas por un facilitador y elegidas de acuerdo al interés del grupo de investigación por la contribución con sus ideas y experiencia. Usualmente las sesiones de grupo son grabadas y posteriormente discutidas, resaltando que inicialmente la tarea no es generar ítems específicos, si no sugerir temas generales, que el grupo de investigación pueda utilizar para elaborar los ítems en sí. Una vez los ítems son escritos, los grupos focales pueden nuevamente juzgar su relevancia, claridad y ambigüedad en términos de la persona que va a responder la escala, y si se han cubierto todos los tópicos de interés.

2.1.2 Entrevistas con informantes claves: Como su nombre lo indica, se realizan entrevistas a un número pequeño de personas, elegidas por su conocimiento y experiencia. Estas personas pueden ser pacientes que tienen o han tenido un trastorno, o se puede realizar también con clínicos que han tenido una experiencia exhaustiva con este tipo de pacientes y que pueden explicar adecuadamente la orientación diagnóstica. Estas entrevistas pueden ser "no estructuradas", las cuales le dan un carácter de espontaneidad a la entrevista o "estructuradas", donde el entrevistador ha planeado cuidadosamente con anterioridad las preguntas.

2.1.3 Observación clínica: Esta técnica es quizás una de las vías más productivas para obtener ítems para una escala específica. Las escalas que provienen de la observación clínica, reúnen en sí la experiencia teórica, práctica e investigativa de sus participantes y sistemáticamente aseguran cierta concordancia entre los observadores con respecto a un trastorno de interés.

2.2 METODOS DE RESPUESTA PARA ESCALAS ESPECIFICAS

Una vez seleccionados los ítems con los métodos mencionados anteriormente, el siguiente paso es

elegir la escala de respuesta según el tipo de pregunta diseñada.

Es importante diferenciar qué tipo de respuestas son de naturaleza categórica como la raza, religión y estado civil, de aquellas variables que poseen una naturaleza continua, como por ejemplo la hemoglobina, presión sanguínea, o la cantidad de dolor registrada en milímetros utilizando una línea recta.

Otra consideración importante para analizar la respuesta de las escalas, es elegir apropiadamente el nivel de medición de las variables. Si la respuesta consiste en categorías tales como género, clasificación laboral o religión, la variable es nominal; si estas categorías poseen un ordenamiento jerárquico como los estadios de un tipo específico de cáncer, nivel educativo, su nivel de medición es ordinal. Algunas variables cuantitativas como la temperatura no refleja un "cero" real en su medición y por esta razón su clasificación es de intervalo, en contraposición con aquellas variables que tienen un "cero" absoluto en su medición como la edad, peso o la talla de un individuo, siendo el caso de las variables medidas a nivel de razón.[7]

Concomitante con la elección apropiada de la naturaleza y nivel de medición de las variables, es necesario tener en cuenta tres problemas que pueden llevar a sesgos en la respuesta de una escala:

El primero está relacionado con la percepción de las respuestas positivas entre diferentes individuos. En segundo lugar, si todos los individuos tienen una percepción muy definida de la respuesta, puede introducirse un error en la medición debido a la dificultad del individuo en encontrar una respuesta apropiada, por la limitación de la escala en su nivel de respuesta. El tercer problema es una consecuencia del segundo y se relaciona con la dicotomización de variables continuas y la pérdida de eficiencia del instrumento así como la reducción de su correlación con otras variables. Sin embargo, Suissa en 1991, demostró como los resultados dicotómicos pueden llegar a tener hasta un 67% de eficiencia con respecto a datos continuos, dependiendo de cómo fue dicotomizada la medición y por esta misma razón, también procedimientos inadecuados pueden disminuir la eficiencia hasta un 10%.

A su vez, es importante anotar que cada vez que se lleve a cabo este procedimiento, es necesario incrementar el tamaño de muestra de la siguiente manera: si se necesitan 67 pacientes para demostrar un efecto cuando el resultado es medido en forma continua, se necesitarán 100 pacientes para demostrar el mismo efecto al llevar a cabo una dicotomización de las variables.[8]

2.3 CATEGORIAS DE RESPUESTA PARA MEDICIONES CONTINUAS

Este tipo de respuestas posee dos categorías: métodos de estimación directa, en la cual se necesita que la persona que responde marque en una línea, la magnitud de su respuesta y técnicas comparativas, en la cual los individuos eligen entre una serie de alternativas que han sido previamente estandarizadas para obtener grupos de criterios independientes.

Las escalas según la forma de respuesta que permiten ambos abordajes son:

- Escala de Guttman
- Escala de Thurstone
- Escala de Likert
- Escala Visual Análoga
- Escala Semántico Diferencial

2.3.1 Los métodos de la estimación directa

Permiten extraer directamente estimaciones cuantitativas acerca de la magnitud del atributo de la persona y su respuesta, la que puede ir desde "totalmente de acuerdo" hasta "totalmente en desacuerdo". En esta clasificación se encuentran las escalas Visual Análoga, la escala de Likert y la escala Semántico Diferencial:[7]

Escala de Likert

Esta escala se usa habitualmente para cuantificar actitudes y conductas. Los individuos que van a

responder el cuestionario reciben una lista de afirmaciones o preguntas y se les indica que seleccionen la respuesta que mejor represente el rango o el grado de su opinión o forma de pensar.

Las opciones de respuesta de esta escala van desde "totalmente de acuerdo" a "totalmente en desacuerdo", puntuadas numéricamente de 0 a 4, o de 1 a 5.

El puntaje total de la escala resulta de la suma de los valores de cada ítem. Posteriormente el autor de la escala debe entrar a definir el rango para las categorías de riesgo.

Escala Visual Análoga

El proceso de construcción consiste en definir el fenómeno, establecer dos valores extremos y opuestos por ejemplo "completamente de acuerdo" y "completamente en desacuerdo" y luego unir esos dos extremos mediante una línea. El sujeto tiene la posibilidad de marcar un punto que se acerque al extremo que él considere pertinente.

El puntaje será la distancia de un extremo de la escala, usualmente 100 milímetros de longitud, al punto señalado por el encuestado. Esta respuesta puede analizarse en forma muy simple considerándola finalmente como "si" y "no", "de acuerdo" y "en desacuerdo". Puesto que cada respuesta contiene un valor numérico que representa una distancia al origen; puede utilizarse además para asignar un puntaje que indique el grado de acuerdo o desacuerdo del encuestado, particularmente cuando se usa en conjunto con otras respuestas que evalúen el mismo constructo.

Este método ha sido usado ampliamente en medicina para evaluar una variedad de constructos como la medición del dolor.[9,10]

Escala Semántico Diferencial

Escala en la cual un sujeto califica cada ítem por medio de una lista de pares de adjetivos, los que deben ser opuestos: "bueno", "malo", unidos por siete a nueve fragmentos de línea. El encuestado debe indicar su adherencia a uno de los adjetivos dados marcando el fragmento de línea que se acerque más a su opinión o concepto.[11]

La escala se puede considerar como una medición en rangos, o sea como una medición de tipo ordinal. Los siete o nueve rangos se deben puntuar y el puntaje total es igual a la suma de los ítems.

2.3.2 Los métodos de la estimación cuantitativa

Esta clasificación comprende las técnicas comparativas o acumulativas de respuesta, la cual consta en las escalas de Guttman y de Thurstone.

Escala de Guttman

También es conocida como escalograma o escala acumulativa. Se trata de un tipo de escalas que tiene como propósito establecer un valor para una variable continua y sólo un atributo subyacente.

El proceso de construcción de una escala de Guttman consiste en definir para el constructo a medir un conjunto de ítems que lo componen y un grupo de expertos determina qué ítems están relacionados con el constructo; los valores de la respuesta son de tipo "si" o "no", a los que se le asignan valores numéricos, siendo 0 la opinión negativa y 1 el caso contrario. El valor final de la escala se obtiene mediante la suma de los valores de cada uno de los ítems.[12]

En la escala, la segunda pregunta cierra un poco el campo de la primera y la tercera a su vez cierra el campo de la segunda y así sucesivamente, de tal forma que es fácil predecir qué respuestas están positivas teniendo como referente cualquier pregunta de la escala.

El puntaje puede ser analizado estadísticamente y corresponde a la suma de los ítems que se respon-

den en forma positiva; cada ítem tiene igual ponderación al resto que compone la escala.

Escala de Thurstone

El método comienza con la selección de 100 a 200 referentes teóricos relevantes del tópico a evaluar. Cada referente es editado en tarjetas separadas y un grupo de expertos los ordena desde el menos acertado hasta el referente más congruente. El siguiente paso es obtener una mediana de cada referente y a cada uno se le asigna el valor correspondiente a la posición que ocupó. Finalmente se seleccionan de 20 a 25 ítems que serán incluidos.[13]

Es una escala de medición en la que las opciones de respuesta son también de "si" y "no" y a diferencia de la escala de Guttman, pueden existir ítems dentro de la escala que tenga un mayor peso que otros.

Una vez los ítems han sido generados teniendo en cuenta los aspectos anteriormente mencionados, debemos asegurar que tan correlacionados están con el constructo o constructos de interés, así como su capacidad discriminatoria, con el fin de evitar incluir sesgos en la medición que lleven a conclusiones erróneas.

3. LA CALIDAD DE UNA MEDICION CON VALIDEZ Y CONFIABILIDAD

El concepto tradicional que ilustra los criterios de validez y confiabilidad es el del practicante de tiro al blanco, el cual debe aprender inicialmente a disparar al centro, esto equivale a hablar en epidemiología de validez.[7]

Este practicante además debe hacer esfuerzos para que cada uno de sus tiros, queden cerca unos de otros, esto es lo que en epidemiología se conoce como la confiabilidad.

El arquero puede hacer disparos repetidos y estos quedar juntos pero lejos del centro. Lo ideal al realizar mediciones en salud es que las dos condiciones vayan juntas. La figura 1 muestra la representación gráfica del criterio de validez y confiabilidad; este último concepto será discutido más adelante.[13]

FIGURA 1. Representación gráfica del criterio de validez y confiabilidad.

3.1 LA VALIDEZ

La validez es comúnmente definida como la capacidad de la prueba para medir lo que se intenta medir, esta definición es muy semejante a la noción de sensibilidad en epidemiología. La forma tradicional de evaluar la validez de una prueba es comparándola con un gold estándar, lo cual deriva los valores clásicos de sensibilidad, especificidad, valores predictivos positivos y negativos.[7]

Con frecuencia en salud mental no es posible disponer de un gold estándar por lo que el proceso de validación va ir apoyado por otros criterios: validez de contenido, validez de criterio, validez de constructo y validez factorial.

3.1.1 Tipos de validez[14]

- **Validez de contenido:** Con el fin de evitar conclusiones no válidas a partir de un test, la información que se obtiene de los ítems, debe satisfacer adecuadamente los criterios conceptuales y el alcance de los mismos para explicar un fenómeno en salud.

  Estos criterios de expertos y la valoración crítica de los ítems que deben ser incluidos en una escala, representan la validez de contenido, la cual

mide el grado de correlación de las preguntas incluidas en la escala y es equivalente al concepto de sensibilidad, en donde se mide la capacidad del test para determinar las personas realmente enfermas cuando este es positivo.

A su vez el concepto clásico de especificidad, que en psicometría corresponde a la validez discriminante, se refiere a la proporción de personas que no poseen la enfermedad y están correctamente diagnosticados.

La validez de contenido se logra idealmente en un estudio que aplique el gold estándar y el test de comparación en toda la población seleccionada. Debido a que esta situación no es normalmente viable, deben seleccionarse personas con y sin la condición de interés, a partir del gold estándar de referencia. Estos pacientes pueden estar hospitalizados o hacer parte de un servicio de tratamiento.

Una de las formas de evaluar estadísticamente la validez de contenido es aplicando las correlaciones de Pearson, cuyos valores negativos indican una correlación inversa, mientras que los positivos indican una correlación directa.

Las correlaciones pueden mostrar asociación entre variables, pero nunca indican acuerdo entre ellas.[3]

- **Validez de criterio:** El segundo tipo de validez es la validez de criterio o predictiva y se refiere a la concordancia empírica del instrumento en estudio, con otras técnicas que miden la misma característica en tres momentos del tiempo: antes, durante y después de que el instrumento es aplicado.

  Un ejemplo de la validez de criterio para trastornos psiquiátricos específicos, puede ser relacionado con la predicción de los resultados. El estudio de Addington y otros en 1993, presentó una serie de ejemplos típicos para usar técnicas correlacionales comparando dos escalas autoaplicadas que miden depresión en pacientes esquizofrénicos: la escala BDI (Beck Depression Inventory) y la escala CDS (Calgary Depression Scale). Los autores evaluaron validez de criterio debido a que el juicio diagnóstico fue llevado a cabo por clínicos entrenados para identificar depresión en población general, a través de dos pruebas para este fin y los resultados fueron analizados utilizando coeficientes de correlación producto-momento de Pearson para identificar la distribución de cada escala, en general y en diferentes subgrupos de pacientes deprimidos.[15]

- **Validez de constructo:** Este tipo de validez involucra procesos más complejos y es esencial para todos los conceptos abstractos como calidad de vida.

  La validez de constructo comienza con la definición conceptual del tópico o constructo que debe ser medido, analizando la estructura interna de sus componentes y la relación teórica de los resultados de la escala con criterios de acuerdo a la evidencia encontrada.

  Robins y Guze en 1972 aportan el primer criterio para evaluar la validez de constructo de los trastornos psiquiátricos al establecer la descripción clínica del trastorno, especificando la sintomatología, factores de riesgo y condiciones precipitantes.[16]

  El segundo criterio se refiere a la relación diagnóstica de la prueba con un test de laboratorio, en donde es importante anotar que actualmente en psiquiatría no existe una prueba serológica o inmunológica, que sirva de gold standard para la validación diagnóstica.

  El tercer criterio para la validación involucra la historia familiar, de esta manera la presencia del mismo trastorno en sus integrantes, puede ser utilizado como indicador de validez del evento.

  El cuarto criterio está relacionado con la validez de predicción de los resultados, los cuales están relacionados con el diagnóstico de interés, incluyendo la respuesta al tratamiento.

Se asume que los individuos con el mismo diagnóstico tendrán resultados similares. Sin embargo el uso de los resultados como criterio de validación es problemático porque muchos trastornos psiquiátricos poseen resultados heterogéneos.

- **Validez factorial:** La validez factorial ha sido utilizada para describir constructos subyacentes en una escala. En este tipo de validez se examina qué tan alejados están los ítems del constructo latente, que pueden no reconocerse fácilmente a través de los datos, es una de las formas más sencillas de evaluar la validez de contenido.

  El análisis factorial ha sido ampliamente utilizado en ciencias sociales, particularmente en la elaboración de escalas psicométricas, permitiendo estudiar la correlación entre un gran número de variables, agrupándolas inicialmente en factores que reflejen el significado de las variables de interés a través de los datos de la investigación.[17]

3.2 LA CONFIABILIDAD

Una medición es muy precisa cuando presenta prácticamente el mismo valor cada vez que se mide, lo cual no significa que sea válida; esta precisión en mediciones sucesivas se conoce como "Confiabilidad".[18]

Cuando se determina el peso corporal a través de una balanza de precisión, la posibilidad de error es menor, en comparación con la medición de calidad de vida a través de un instrumento diseñado para este fin; en este último, la oportunidad de generar valores distintos en dos ocasiones sucesivas es mayor.

Al efectuar una medición existen tres fuentes principales de error:

- La variabilidad del observador, lo cual se refiere a la persona que realiza la medición, e incluye factores como la elección de las palabras al realizar una entrevista o en la elaboración de ítems de una escala.

- Del observado, la cual hace referencia a la variabilidad biológica intrínseca de los sujetos que se estudian, debida a fenómenos como los ritmos circadianos y el tiempo transcurrido en la última toma de una medicación.

- Finalmente la debida al instrumento, que representa la variabilidad de la medición por una inadecuada calibración del mismo, cuando este instrumento es mecánico, o por falta de estandarización de los entrevistadores cuando se aplica una encuesta.

3.2.1 Estimación de la confiabilidad

La precisión de una variable determinada se describe frecuentemente en estadística básica a través de la desviación estándar o el coeficiente de variación de una serie de mediciones sucesivas; sin embargo cuando se utilizan escalas, la precisión puede valorarse examinando la consistencia entre los resultados y su concordancia aplicando coeficientes de correlación.

Los tres requisitos para valorar la confiabilidad de las mediciones son:[19]

- **Consistencia interna**: Define el grado de concordancia entre dos variables que miden la misma característica general.

- **Consistencia test-retest**: Es el grado de concordancia entre mediciones repetidas en una muestra de individuos. Debe seleccionarse muy cuidadosamente el intervalo de tiempo, ya que si es demasiado largo, la falta de concordancia entre los resultados puede deberse a variaciones que tengan un significado propio (no debidas al azar), mientras que si es muy corto es posible que no haya tiempo suficiente para que aparezcan las fluctuaciones debidas al azar.

- **Consistencia ínter e intraobservador**: La consistencia interobservador es la correlación entre los valores obtenidos por dos observadores o más, en una misma muestra de individuos, mientras que la consistencia intraobservador es la correlación entre valores repetidos obtenidos por un mismo observador.

BIBLIOGRAFIA

1. Corcoran K, Fisher J. Measures for clinical practice: a sourcebok. New York. Free Press. 1987.
2. Katz S, Akpon CA, et al; Measuring the health status of populations. En: Berg RL, editor. Health Status Indexes. Chicago: Hospital Research and Educations Trust. 1973. p. 39-52.
3. McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaries. 2a edición. New York: Oxford University Press; 1996. p. 10-15.
4. Bombardier C, Tugwell P. A Methodological framework to develop and select indices for clinical trials: statistical and judgmental approaches. J Rheumatol 1982; 9: 753-7.
5. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to their Development and Use. Second edition. New York: Oxford University Press; 1994. p. 162-185.
6. Willms DG, Johnson NA. Essentials in qualitative research: A notebook for the field. Unpublished manuscript. 1993.
7. Trochim W. General Issues in scaling. Junio 2002. URL disponible en: http://trochim.omni.cornell.edu/kb/scalgen.htm.
8. Suissa S. Binary methods for continuous outcomes: A parametric alternative. Journal of Clinical Epidemiology 1991; 44: 241-8.
9. Huskisson EC. Measurement of pain. Lancet November 9 1974; 1127-31.
10. Scott PJ, and Huskisson EC. Measurement of functional capacity with visual analogue scales. Reumathology and Rehabilitation 1978; 16: 257-9.
11. Osgood C, Suci G, Tannenbaum P. The measurement of feeling. University of Illinois Press: Urbana; 1957.
12. Trochim W. Guttman scaling. Junio 2002. URL disponible en: http://trochim.omni.cornell.edu/kb/scalgutt.htm.
13. Trochim W. Introduction to validity. Junio 2002. URL disponible en: http://trochim.omni.cornell.edu/kb/introval.htm.
14. Trochim W. Reliability and Validity: What's the difference?. Junio 2002. URL disponible en: htpp://trochim.human.cornell.edu./tutorial/colosi2.htm.
15. Addington D, Addington J, Maticka-Tyndale E. Rating depression in Schizophrenia: A comparison of a self-report and an observer scale. J Nerv Men Dis 1993; 181:561-5.
16. Robins E, Guze SB. Establishment of diagnostic validity in psychiatric illness: its application to schizophrenia: Am J Psychiatry 1970; 126(7): 983-7.
17. Afifi AA, Clark V. Computer-Aided Multivariate Analysis. New York: Van Nostrand Reinhold; 1990.
18. Trochim W. Reliability. Junio 2002. URL disponible en: http://trochim.omni.cornell.edu/kb/reliable.htm.
19. Trochim W. Types of reliability. Junio 2002. URL disponible en: http://trochim.omni.cornell.edu/kb/reltypes.htm.