Rev. Salud pública. 6 (3): 302-318, 2004

GUIA/GUIDE

# Validación de Escalas de Medición en Salud

RICARDO SÁNCHEZ[1] y JAIRO ECHEVERRY[2]

[1] Médico. Especialista en Psiquiatría. Especialista en Estadística. M. Sc. Epidemiología Clínica.. Facultad de Medicina, Universidad Nacional de Colombia.
E-mail: rsanchezpe@unal.edu.co
[2] Médico. Especialista en Pediatría. M. Sc. (candidato) Epidemiología Clínica. Facultad de Medicina, Universidad Nacional de Colombia.
Correo electrónico: jecheverr@unal.edu.co

**Recibido 30 Junio 2004/Enviado para Modificación 12 Julio 2004/Aceptado 26 Agosto 2004**

## RESUMEN

Este artículo muestra los aspectos más importantes relacionados con la validación de escalas aplicadas en investigación en salud e ilustra los pasos de este proceso. En el artículo se discuten los conceptos relacionados con la selección de los ítems, traducción, validez, confiabilidad y utilidad.

**Palabras Claves**: Estudios de validación, reproducibilidad de resultados (*fuente: DeCS, BIREME*).

## ABSTRACT

**Validating scales used for measuring factors in medicine**

This article shows the mayor issues related with validation of scales applied in health research and illustrates the steps of this process. Concepts related with the selection of items, translation, validity, reliability and usefulness are discussed in this article.

**Key Words**: Validation studies, reproducibility of Results (*source: DeCS, BIREME*).

El método científico, en su positivismo lógico, se ve avocado a un reduccionismo necesario pero a veces limitante, que supone colapsar conceptos anatómicos, fisiológicos, fisiopatológicos, nosológicos, psicológicos, antropológicos o sociales, en variables que, como sus representantes, deben ser medidas (1). El establecer la frecuencia de un evento relacionado con la salud, o determinar los factores que con él se asocian, implica una adecuada selección de los seres humanos que serán objeto del estudio y una cuidadosa medición de características que denominamos variables (2). Los aciertos o errores en el proceso de medición de estas variables pueden afectar la validez de los datos: esto quiere decir que se puede perder la correspondencia entre el resultado de la medición y la realidad del fenó

meno que se está estudiando. Por lo tanto, medición y validez son dos elementos inseparables en toda investigación científica.

De manera elemental la medición se define como situar algo al lado de un patrón de referencia para ver a qué valor de ese patrón corresponde. Por ejemplo, existe un patrón de referencia que define al sexo como masculino o femenino, siendo masculino quien posea un patrón cromosómico XY y femenino quien lo posea XX. En este caso el patrón de referencia tiene sólo dos niveles de medición que se denominan categorías o modalidades.

Si un individuo tiene un patrón cromosómico XY, queda clasificado o medido como masculino de acuerdo con ese patrón de referencia. Pero sabemos que en la práctica para clasificar o medir el sexo no es necesario hacer un estudio cromosómico, ni efectuar un exhaustivo análisis de los genitales externos, ya que simplemente con el nombre del individuo podemos aproximarnos bastante a esta medición: esto sucede porque, para medir el sexo, en el necesario proceso reduccionista a que nos hemos referido, los dos patrones de referencia -estudio cromosómico y nombre- tienen un alto nivel de concordancia.

Existen otros patrones de referencia que tienen un nivel muchísimo mayor de categorías: por ejemplo una cinta métrica de 2 metros de longitud marcada cada centímetro, tendrá 200 categorías. Si utilizamos este patrón de referencia para medir la estatura y colocamos a un individuo al lado de este patrón, podríamos encontrar que el individuo cae en la categoría 175: en este caso diríamos que el individuo mide 175 centímetros. Aquí pierde sentido el hablar de categorías o modalidades y es más adecuado referirse a intensidad o cantidad de algún atributo (en este caso la estatura) (3).

En el anterior ejemplo se han definido dos estrategias de medición:

a. Una que permite clasificar o categorizar a los individuos o unidades de medición, cualifica un atributo.
b. Otra que cuantifica un atributo.

De manera general puede decirse que los patrones de referencia se pueden concretar a través de los instrumentos de medición, de la misma manera que un concepto o cualidad se concreta en una variable. Dependiendo del diseño del instrumento, puede obtenerse una mejor calidad de la medición: por ejemplo, una cinta métrica marcada en milímetros proporciona una medida más adecuada y real que una marcada en centímetros. En este caso será un mejor instrumento de medición para cuantificar ese atributo o realidad (4).

## LA MEDICIÓN POR MEDIO DE ESCALAS

A veces la condición que se quiere medir no se puede delimitar de manera precisa, o no puede establecerse para ella una medida simple: este caso es frecuente cuando el fenómeno de interés es muy complejo y presenta características en diferentes niveles, tal y como ocurre usualmente en fenómenos psicológicos o sociales, donde muchos de los diagnósticos o definiciones que usualmente se manejan corresponden a categorías más bien inespecíficas.

Un ejemplo que permite ilustrar esta situación es el síndrome depresivo: asumiendo que este realmente existe, debemos reconocer que no es posible desarrollar un instrumento que lo mida con precisión ya que el síndrome mismo está imprecisamente definido. Esto equivale a decir que no existe un verdadero patrón de oro para su determinación, y que más bien éste podría ser construido en función de la elaboración conceptual que haga, para cada condición en particular, un grupo de expertos en un marco cultural específico.

Sabemos que no en todos los pacientes deprimidos hay tristeza pues en algunos hay anhedonia ( incapacidad para experimentar placer), que algunos pacientes presentan anorexia mientras que otros tienen hiperfagia, que no todos los pacientes tienen ideas de culpa, o que algunos pacientes están motoramente inhibidos mientras que otros están agitados, por citar solo algunas características del trastorno. Esto implica que la realidad que estamos tratando de evaluar no puede ser medida ni observada directamente. Para efectuar la medición en estos casos se recurre a una estrategia que es agrupar las características de la entidad en categorías un poco más gruesas, que siempre están presentes. En el ejemplo anterior podríamos descomponer el síndrome depresivo en las siguientes categorías:

a. Alteraciones del estado de ánimo.
b. Alteraciones de la conducta motora.
c. Alteraciones de funciones autónomas.
d. Alteraciones del pensamiento y de la función cognoscitiva.

De esta manera, así el paciente esté triste o anhedónico tendrá alteraciones del estado de ánimo; independientemente de si está hipoquinético o agitado, tendrá alteraciones en la conducta motora; lo mismo sucedería en las otras categorías en las cuales se ha descompuesto el síndrome. El anterior procedimiento es lo que se conoce como desarrollo de una estructura de Dominios o Factores.

El siguiente paso en la construcción de escalas es medir cada uno de los Dominios o Factores, lo cual habitualmente se hace mediante preguntas o aspectos de exploración específicos (ítem) cuya respuesta o resultado puede darse en un escalamiento categórico o continuo.

La resultante colección de ítem que miden Factores, que a su vez constituyen un síndrome que no es directamente mensurable, es lo que se denomina una escala (5). Si se toma el listado de preguntas de la última columna de la Tabla 1, se tiene una escala que en este caso ha sido diseñada para la medición de una realidad que denominamos Síndrome Depresivo (6).

El ejemplo presentado ilustra el caso de una escala diseñada para incluir un individuo en una categoría. En otros casos las escalas buscan establecer la intensidad de un atributo, lo cual es de utilidad cuando se desea establecer la magnitud de una característica a lo largo del tiempo (7).

**Tabla 1**. Estructura interna de una escala

| Sindrome (Concepto) | Factores (Dominios) | Ítem |
|---|---|---|
| Síndrome Depresivo | Alteración del estado de ánimo | ¿Se ha sentido triste?<br>¿Se ha sentido aburrido?<br>¿Ha perdido capacidad de disfrutar? |
| | Alteración de la conducta motora | ¿Ha disminuido la actividad motora?<br>¿Se fatiga con facilidad?<br>¿Presenta agitación motora? |
| | Alteración de funciones autónomas | ¿Ha tenido insomnio?<br>¿Ha tenido anorexia?<br>¿Ha aumentado el apetito? |
| | Alteración de pensamiento y función cognoscitiva | ¿Se siente culpable de algo?<br>¿Ha pensado en suicidarse?<br>¿Tiene dificultad para concentrarse? |

## ¿EN QUÉ CONSISTE Y CÓMO SE REALIZA LA VALIDACIÓN DE UNA ESCALA?

El proceso de construcción de una escala puede esquematizarse en la siguiente secuencia:

a. Estamos ante algo que creemos que realmente existe y que queremos medir.

b. Lo que queremos medir es un fenómeno complejo, que no es observable directamente y que tiene manifestaciones o maneras de expresarse que no son constantes.
c. Se descompone lo que queremos medir en grupos de síntomas o manifestaciones que siempre están presentes. Esos grupos se llaman Factores o Dominios.
d. Se desarrolla una estrategia para medir los Factores o Dominios, usualmente mediante la construcción de ítem.

Pero este proceso no se detiene aquí: el siguiente paso es "certificar" que la escala tiene ciertas características o atributos que hacen meritoria su utilización. Esas características, de las cuales depende la "certificación", son:

a. La realidad existente está adecuadamente representada por la escala. En otras palabras, la escala parece medir lo que debe medir.
b. La escala refleja la estructura de Dominios o de Factores en los cuales fue dividida le realidad que se va a medir. En otras palabras, la escala no deja Factores sin medir, ni mide Dominios que no corresponden a la realidad.
c. La escala funciona de manera parecida a otros instrumentos para medir esa realidad que ya han sido "certificados".
d. La escala funciona bien bajo diferentes condiciones, por ejemplo cuando se aplica en diferentes oportunidades o cuando es aplicada por distintas personas.
e. Cuando la realidad que se está midiendo cambia, la escala puede detectar ese cambio.
f. La escala es práctica, fácil de aplicar y de procesar.

El proceso de "certificación" que se ha presentado es lo que se denomina "validación de escalas". Utilizando una terminología técnica, los pasos anteriores pueden describirse así:

a. Validez de Apariencia: La escala parece medir lo que debe medir.
b. Validez de Constructo: La escala no deja factores sin medir ni mide dominios que no son del síndrome.
c. Validez de Criterio: La escala funciona parecido a otros instrumentos.
d. Confiabilidad test–retest o interevaluador: La escala funciona bien bajo diferentes condiciones.
e. Sensibilidad al Cambio: La escala detecta modificaciones de la realidad que mide.
f. Utilidad: Es una escala fácil de aplicar y procesar.

En el área de la salud existen muchísimos instrumentos de medición que caben en el concepto de escala. Es lógico que esto sea así, dada la complejidad de muchos de los fenómenos que se tienen que medir (actitudes, creencias, comportamientos, adaptación, autonomía, ajuste social, etc.). La mayoría de estas escalas ha sido desarrollada en países de habla inglesa, lo cual genera una serie de inconvenientes cuando se van a utilizar en países con lenguajes o culturas diferentes. El tener que aplicar estas escalas en pacientes de una cultura diferente de la cultura en la cual se desarrolló, por parte de clínicos de una cultura distinta de la de los clínicos con quienes se creó el instrumento y con un ajuste lingüístico que supone la traducción, implica que se debe repetir el proceso de certificación del instrumento, es decir, de validación de la escala (8).

## ¿POR QUÉ VALIDAR ESCALAS?

De acuerdo con los estándares sobre selección, construcción y aplicación de pruebas psicológicas, existen las siguientes recomendaciones (9):

1. Cuando se hacen cambios sustanciales en el formato del instrumento, modo de aplicación, idioma o contenido, el usuario debería revalidar la escala para las nuevas condiciones, o tener argumentos que apoyen que no es necesaria o posible una validación adicional.
2. Cuando se traduce una escala de un idioma o dialecto a otro, debe establecerse su fiabilidad y validez en los nuevos grupos lingüísticos en los que se aplique.
3. Cuando se pretende que las dos versiones de una escala en idiomas distintos sean comparables, hay que aportar pruebas de la comparabilidad de los instrumentos.

Dadas las anteriores recomendaciones, podría optarse por realizar una nueva escala en lugar de emprender una validación. Sin embargo existe una serie de argumentos que están a favor de la validación:

1. Puede resultar más económico y rápido, o sea más eficiente, hacer una validación que desarrollar un nuevo instrumento.
2. Al utilizarse instrumentos mundialmente aceptados se abre la posibilidad de efectuar estudios entre diferentes países o entre diferentes culturas.
3. Las escalas con amplia aplicación clínica suelen ser instrumentos suficientemente probados.
4. El desarrollo de una nueva escala implica disponer de recursos técnicos y humanos altamente calificados y con experiencia en el área.

Las razones para validar escalas parecen ser claras, no así los métodos para hacerlo (10). Existen numerosas prácticas que se han tomado como validaciones sin serlo: las más frecuentes son efectuar únicamente traducciones, realizar solamente pruebas de concordancia o correlación con los resultados de la medición de otro instrumento, o practicar tan solo pruebas de concordancia entre diferentes evaluadores. Se debe ser cuidadoso al utilizar un instrumento validado para fines clínicos o de investigación, dado el peligro de efectuar mediciones que no coincidan con la realidad.

## ¿CUÁL ES EL PROCESO DE VALIDACIÓN DE UNA ESCALA?

La validación de una escala es un proceso complejo, que comprende varias etapas y que requiere un número elevado de pacientes. Supone además disponer de herramientas estadísticas para efectuar procedimientos que pueden resultar complejos. En general, el proceso de validación de una escala debe tener las siguientes etapas:

Selección de la Escala
Traducción
Pruebas Preliminares para realizar ajustes (de ítems y utilidad)
Pruebas de su Validez
Pruebas de su Confiabilidad
Determinación de su Utilidad

### 1. Selección de la escala

Aunque este paso parezca obvio, es importante resaltar su importancia. La escala que se va a validar debe ser la mejor disponible. Saber esto implica efectuar una cuidadosa revisión sistemática de la literatura disponible sobre el tema y eventualmente efectuar una consulta a expertos en el área sobre la que se efectuará la medición. En el reporte de validación de una escala debe quedar suficientemente fundamentado por qué se seleccionó esa escala para ser validada.

### 2. Traducción

Cuando la escala original se encuentre en otro idioma, debe someterse a un proceso de traducción. Este proceso debe hacerse con cuidado ya que puede introducir distorsiones en la escala y por lo tanto generar errores de medición (11).

El énfasis del proceso no es la traducción textual o literal, sino la traducción del sentido conceptual que cada ítem persigue.

El primer paso en el proceso de traducción es conformar un comité de revisión; este comité estará compuesto por miembros del grupo encargado de la revisión (al menos dos personas), por una o dos personas con experiencia en el área de aplicación del instrumento y conocimiento del idioma original de la escala y por uno o dos sujetos que hayan presentado o que presenten la condición que busca medir el instrumento. Antes de buscar traductores, el comité de revisión debe prever dificultades que puedan surgir en la traducción o validación: por ejemplo, si en una escala que se va a aplicar en pacientes colombianos se incluye un ítem que explora la afición por el jockey sobre hielo, se puede sugerir que se explore mejor la afición por el fútbol. Una vez definido este punto, el instrumento es enviado a los traductores. El número de traductores requerido no es tan importante como la calidad de los mismos. En general se recomienda buscar traductores con las siguientes características (12):

a. Ser competentes en los idiomas implicados (bilingüismo).
b. Conocer, ser parte o estar inmerso en la cultura en la cual se aplicará la escala validada.
c. Tener un entrenamiento básico sobre medición en salud o comportamiento.
d. Tener un mínimo entrenamiento en construcción de instrumentos de medición y *"tests"*.

Para cumplir las dos últimas características se puede efectuar una capacitación sobre estos temas a los traductores que participen en el proceso. La traducción del idioma original al idioma de aplicación se denomina traducción directa; la traducción en el otro sentido se llama traducción inversa.

Existen diferentes estrategias para garantizar una traducción de la mejor calidad (13). Una de las más prácticas y sencillas recomienda disponer de por lo menos dos traducciones directas efectuadas por diferentes individuos (traductor A y traductor B). Si se encuentran diferencias importantes en el significado de algunos de los ítems, debe solicitarse a los traductores que en conjunto, lleguen a un acuerdo sobre los puntos discordantes. Si esto no es posible debe recurrirse a un tercer traductor. Dependiendo de las características del instrumento, después de un tiempo en el cual se garantice que los traductores no recuerden el contenido de la escala que tradujeron previamente (al menos una semana), se entrega la versión del traductor A al B y la del B al A para que efectúen una traducción inversa. Entonces el comité de

revisión evaluará si el significado de los ítems re-traducidos coincide con los de la escala original.

Otra alternativa es entregar la versión traducida a un tercer evaluador que hasta ahora no haya participado en el proceso y que se encargará de efectuar la traducción inversa. Con estos insumos el comité de revisión establecerá una versión traducida del instrumento.

3. Prueba preliminar para ajuste

La versión traducida que se definió en la fase previa se aplicará a un grupo de pacientes (entre 10 y 15) por parte de dos o tres evaluadores diferentes. Tanto los pacientes como los evaluadores deberán tener características similares a las del escenario de aplicación final de la escala. En esta fase se analizan aspectos relacionados con particularidades de los rubros y con la utilidad de la escala (14).

a. En relación con las particularidades de los rubros se debe tener en cuenta:

- Grado de comprensión de los rubros: Como las escalas buscan medir condiciones presentes independientemente del nivel educativo de los pacientes, los rubros deben estar escritos en un lenguaje sencillo, fácilmente comprensible para todos, evitando términos técnicos o de uso poco frecuente.
- Ambigüedad: Deben evitarse términos que puedan ser interpretados de distintas maneras. Por ejemplo, en la pregunta: "¿se ha sentido usted deprimido últimamente?", el término "últimamente" puede interpretarse de diferentes maneras (en el último año, en el último mes, después de algún acontecimiento particular, etc) lo cual genera errores en la medición.
Ítem con carga afectiva: La utilización de ciertos adjetivos puede generar errores de medición al inducir algún tipo particular de respuesta. Por ejemplo: "¿Se siente molesto por las *ridículas* equivocaciones de su médico?".
- Frecuencia de respuesta: Si más del 95% de los individuos a quienes se aplica la prueba preliminar para ajuste califican igual un ítem, debe considerarse la posibilidad de excluirlo de la escala, dada la posibilidad de que no esté aportando variabilidad al instrumento. Dependiendo de las características de la escala, una conducta adecuada es poner este ítem "bajo observación" y analizar posteriormente las escasas respuestas no coincidentes, dada la posibilidad que correspondan a errores en la medición.

- Restricción de Rango de Respuesta: Si un ítem puede responderse dentro de las cuatro opciones “bueno–regular–malo–muy malo”, y ninguno de los individuos utiliza la opción “muy malo”, puede pensarse en eliminar esta opción de medición del ítem y dejarlo solo con tres niveles.

b. En relación con la utilidad de la escala deben evaluarse aspectos como:

- El tiempo de diligenciamiento y aplicación requerido: Entre menos tiempo se demore en efectuar la medición bien hecha, mejor.
- Necesidad de entrenamiento: Entre menos entrenamiento especial se requiera para aplicar la escala ésta será más fácil de aplicar en condiciones reales.
- Características del formato del instrumento: Entre más corto el instrumento es más fácil de tramitar. En los textos largos se recomiendan tipos de letra que permitan una fácil lectura, como los que tienen *serifs*. Puede ser necesario incluir un corto instructivo al inicio de la escala.
- Facilidad para calificar el puntaje final de la escala: Los ítem que tienen diferente ponderación o los procedimientos que suponen algoritmos complicados para calificar la escala deberían evitarse.

4. Pruebas de Validez

La evaluación de la validez de una escala busca responder a las siguientes preguntas:

a. ¿La escala parece medir lo que debe medir? *Validez de apariencia*:
Responder a esta pregunta tiene importancia para determinar la aceptabilidad que puede tener la escala en el escenario de aplicación (15). Para establecer la validez de apariencia se deben conformar dos grupos, uno de sujetos que van a ser medidos con la escala y otro de expertos: ellos analizan la escala y dictaminan si ésta realmente parece medir lo que se propone. Esta validez no supone un concepto estadístico, sino que depende de los juicios que los expertos hagan sobre la pertinencia de los ítems. Cada uno de los grupos puede estar conformado por cuatro o cinco personas.
b. ¿Refleja la estructura de dominios del síndrome? *Validez de Contenido*:
La respuesta de esta pregunta supone evaluar si los diferentes ítems incluidos en el instrumento representan adecuadamente los dominios o factores del concepto que se pretende medir. El procedimiento para evaluar la validez de contenido supone aplicar métodos estadísticos como el análisis factorial (16). La ventaja de estos métodos es que permiten saber, no solo cuál es la estructura factorial, sino cómo representan los ítems los distintos factores, y eventualmente retirar ítem que no aportan variabilidad a la medición del sín-

drome. Para efectuar este tipo de análisis se requieren por lo menos cinco pacientes por cada ítem que tenga el instrumento, pero no menos de 100 pacientes en total (17).

Todos los pacientes deben presentar la condición que la escala pretende medir, incluyendo los diferentes espectros de intensidad. Una vez determinados cuáles son los diferentes factores que mide la escala, puede recurrirse a otros instrumentos que midan esos factores para efectuar una comparación: a esto se le denomina *Validez de Constructo*. Por ejemplo, si el análisis factorial en una validación de una escala muestra un dominio denominado Depresión, puede aplicarse simultáneamente una escala reconocida para medir depresión y contrastar su resultado con el de la escala que se está validando. Otra estrategia reconocida para medir la validez de constructo es mediante la evaluación de los valores de correlación en estructuras matriciales (Matriz multirrasgo – multimétodo (18)).

c. ¿Funciona de manera similar a otros instrumentos certificados? *Validez de criterio (concurrente y predictiva)* (19):
Para saber esto debe compararse la escala que se está validando, con un patrón de oro que debería ser la mejor escala disponible en el área de aplicación clínica: en este caso se habla de *validez de criterio concurrente*. Estadísticamente la comparación se efectúa mediante coeficientes de correlación de *Pearson* o de *Spearman*, dependiendo de las características de distribución de los datos. Por supuesto, la comparación debe hacerse con un instrumento ya validado. En caso de que no haya más instrumentos validados la comparación suele efectuarse con métodos de apreciación clínica subjetiva (como la impresión clínica global), pero reconociendo que los valores de correlación con este tipo de instrumentos no suelen ser muy altos, lo cual no quiere decir que la escala funcione mal. Para obviar este inconveniente, una alternativa es efectuar validación simultánea de dos instrumentos que evalúen la misma condición. Esta alternativa tiene la ventaja de que reporta valores de correlación más consistentes y que permite aprovechar una misma muestra de pacientes para el procedimiento de validación. Los valores de correlación deben estar preferiblemente por encima de 0.8. El hallazgo de valores altos de correlación entre las escalas en proceso paralelo de validación debe interpretarse con cautela.

Cuando quiera que se evalúe la correlación o concordancia entre el resultado actual del instrumento y un evento relacionado que puede suceder en el futuro, el procedimiento recibe el nombre de *validez de criterio predictivo (validez predictiva)*. Un ejemplo es la evaluación de un estudiante universitario con"criterios ideales de excelencia" antes de iniciar su entrenamiento y

compararlo con las notas que ese estudiante vaya a obtener en el futuro o al finalizar su carrera.

d. Cuando la condición que se está midiendo cambia, puede la escala medir ese cambio?: *Sensibilidad al cambio*.
Medir la sensibilidad al cambio es de particular importancia cuando se trata de instrumentos diseñados, no tanto para diagnosticar, sino para cuantificar atributos, lo cual nos asegura que la escala es buena para medir una condición a lo largo del tiempo. Este tipo de escalas son las que nos permiten evaluar la respuesta a un tratamiento. El procedimiento más usado para evaluar la sensibilidad al cambio consiste en comparar una puntuación inicial con una puntuación posterior, en un momento en el cual se haya modificado la condición clínica (20). La documentación de este cambio suele hacerse aplicando otras escalas o simplemente una evaluación clínica global. Los métodos estadísticos empleados dependen de las características distribucionales de los puntajes de la escala, aunque usualmente son útiles los Análisis de Varianza para Mediciones Repetidas.

5. Pruebas de confiabilidad

Como se mencionó antes, la confiabilidad hace referencia a si la escala funciona de manera similar bajo diferentes condiciones, dependientes del mismo instrumento, del tiempo de aplicación y del clínico que hace la medición. Se puede decir que la confiabilidad es una medición del error que puede generar un instrumento al ser inestable y aplicarse en diferentes condiciones. Debe evaluarse la confiabilidad en tres aspectos:

a. Relacionados con el instrumento:
Si los ítems que conforman la escala, tienen unos adecuados niveles de correlación entre ellos, conforman una estructura “aglutinada” que le confiere cierta estabilidad al instrumento. Las correlaciones entre ítem con ítem, entre ítem y factor y entre ítem y escala son una especie de pegante que le confiere al instrumento lo que se denomina *consistencia Interna* u *homogeneidad*. La medición de esta consistencia se realiza mediante diferentes procedimientos pero los más usados son el coeficiente KR-20 (fórmula 20 de Kuder–Richardson) y especialmente el alfa de Cronbach (21). El primero de estos instrumentos se usa cuando los ítems son de respuesta dicotómica. El alfa de Cronbach permite evaluar homogeneidad en escalas cuyos ítems pueden responderse en más de dos alternativas.

Al evaluar los resultados de estos coeficientes debe tenerse en cuenta que sus valores se afectan por el número de ítem en la escala; según esto, al aumentar el número de ítem del instrumento el valor del coeficiente alfa se incrementa artificialmente. Los valores que se recomiendan para estos índices son entre 0.7 y 0.9 (70% a 90%). Valores bajos sugieren que la escala es poco homogénea, que puede estar evaluando diferentes fenómenos y que no muestra consistencia ante diferentes condiciones de aplicación; valores mayores de 0.9 sugieren una estructura demasiado homogénea, en la cual probablemente existan ítem redundantes. Los diferentes programas estadísticos existentes, como SPSS®, SAS®, NCSS® y STATA®, por citar algunos, efectúan el cálculo de estos coeficientes.

b. Relacionados con el tiempo de aplicación:
Se debe medir si la escala, cuando se aplica en diferentes momentos, permaneciendo estable la condición que se mide, mantiene un resultado similar en la medición. Esto es lo que se ha denominado *confiabilidad test – retest*. Para medir este tipo de confiabilidad se aplica la escala por lo menos dos veces, en diferentes momentos, en situaciones de estabilidad del fenómeno o síndrome que se está midiendo. La medición de este tipo de confiabildad se ha efectuado mediante diferentes procedimientos:

- Coeficiente de correlación de Pearson: Evalúa cómo se relacionan los puntajes de los diferentes momentos, en términos de asociación lineal. Es un método poco utilizado ya que no incorpora en el análisis otras fuentes de variabilidad, al asumir que toda la varianza es explicada por las diferencias entre los sujetos medidos.

- Coeficiente de correlación intraclase: Es una medida de confiabilidad mejor que la anterior ya que incorpora en el análisis, además de la variabilidad entre los sujetos, otras fuentes de variabilidad como pueden ser diferentes observadores, características de los pacientes (también llamada variabilidad dentro de los sujetos) y error. Se puede calcular a través de un procedimiento estadístico denominado Análisis de Varianza de Mediciones Repetidas (ANOVA de medidas repetidas). El resultado del coeficiente se interpreta como el porcentaje de la variabilidad de los puntajes que depende solo se la variabilidad entre los sujetos medidos. Por ejemplo, si el valor es 0.9 esto indica que el 90% de la varianza de los puntajes depende solo de la variabilidad de los sujetos. Una adecuada confiabilidad test – retest está indicada por valores mayores de 0.8.

- Coeficiente de correlación–concordancia de Lin (22): Este coeficiente se basa en la premisa de que el caso ideal de correlación se da cuando el disper-

sograma que originan dos mediciones se ve como una recta con una inclinación de 45 grados. Los coeficientes tradicionales no pueden detectar esta situación ideal, por lo cual, así reporten valores elevados, no necesariamente están reflejando la concordancia entre las dos mediciones. El coeficiente de Lin compara el acuerdo entre dos pares de mediciones midiendo la variación alrededor de una línea de 45º que parte del origen. Se recomienda usar esta medida como complemento del coeficiente de correlación intraclase.

c. Relacionados con la aplicación por diferentes personas:
Si en el mismo momento, ante el mismo paciente, la escala es aplicada por diferentes observadores, los resultados de la medición deberían ser similares. Esto es lo que se mide con la Confiabilidad interevaluador. Obviamente, los evaluadores deberán tener un entrenamiento similar o una capacitación uniforme para aplicación el instrumento. De otra manera, los puntajes diferentes estarán reflejando, no debilidades de la escala, sino una fuente de variabilidad adicional introducida por quienes efectúan la medición.

El tamaño de muestra requerido para la estimación de estos diferentes coeficientes dependerá del valor estimado de correlación, del número de observadores y del nivel de significación establecido (23).

6. Determinación de la utilidad

Este punto hace referencia a la aplicabilidad del instrumento en el escenario real. Si bien, no depende de la aplicación de procedimientos estadísticos, en la validación de una escala debe describirse el tiempo promedio requerido para aplicar el instrumento, la necesidad de condiciones particulares en las cuales haya que poner al sujeto antes de iniciar el procedimiento, el grado de capacitación o calificación profesional que requieren quienes se encargarán de aplicar el instrumento, y la forma, método y tiempo requerido para calificar el puntaje de la escala.

## ¿CÓMO LEER CRÍTICAMENTE UN ARTÍCULO RELACIONADO CON VALIDACIÓN DE ESCALAS?

Como cada vez hay una mayor tendencia a efectuar validaciones de escalas en nuestro medio, presentamos la siguiente lista de chequeo para facilitar el proceso de análisis y evaluación de los artículos y publicaciones relacionadas con estos procedimientos:

Lista de chequeo para evaluar la calidad de la validación de una escala:

- Item

¿De dónde vienen?:

• De escalas previas.
• De observación clínica.
• De opinión de expertos.
• De reporte de pacientes.
• De hallazgos de investigación.
• De supuestos teóricos.

Se midió en ellos:

• Frecuencia de respuesta
• Restricción en el rango
• Comprensión
• Ambigüedad
• Carga afectiva

- Validez:

- ¿Tiene validez de apariencia?
- ¿Hay evidencia de la validez de contenido?
- ¿Hay evidencia de la validez de constructo?
- ¿Se evaluó validez de criterio? ¿De manera concurrente? ¿De manera predictiva?

- ¿En qué grupos se evaluó la validez?

- Confiabilidad:

– ¿Se evaluó consistencia interna?
– ¿Tiene buena confiabilidad test-retest?
– ¿Y confiabilidad interevaluador?
– ¿En qué grupos se midió confiabilidad?
– ¿Cómo se calculó la confiabilidad?

- Utilidad:

– ¿Se puede aplicar en un tiempo razonable?
– ¿Qué tanto entrenamiento se requiere?
– ¿Es fácil de calificar?

## REFERENCIAS


1. Leach DC. Changing education to improve patient care. Quality in Health Care 2001; 10 (supp II):ii54-ii58.
2. Echeverry J. Definiendo Enfermedad. Rev Col Neumol 2003;15(2):69-81
3. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research. Principles and Quantitative Methods. New York: Van Nostrand Reinhold; 1982.
4. McDowell I, Newell C. The Theoretical and technical Foundations of Health Measurement. In: Measuring Health. 2nd edition. New York: Oxford University Press; 1996. p. 10-46.
5. DeVellis RF. Scale Development. Theory and Applications. Newbury Park: Sage Publications; 1991
6. Sánchez R, Gómez C. Conceptos básicos sobre validación de escalas. Revista Colombiana de Psiquiatría. 1998; 27:121-130.
7. Guyatt GH, Kirshner B, Jaeschke R. Measuring Health Status: GAT is the Necessary Measurement Properties? J Clin Epidemiol. 1992; 45:1341-1345.
8. Guillemin F, Bombardier C, Beaton D. Cross – Cultural Adaptation of Life Measures: Literature Review and Proposed Guidelines. J Clin Epidemiol. 1993;46:1417-1432.
9. American Psychological Association. Standards for Educational and Psychological tests. Washington: Author; 1985.
10. Hambleton RK. Guidelines for adapting educational and psychological tests: A progress report. Eur J Psych Assess 1994; 10: 229-240.
11. Berkanovic E. The Effect of Inadecuate Language Translation on Hispanic Responses on Health Surveys. Am J of Public Health 1980;70:1273-1276.
12. Hambleton RK. Adaptación de Tests para Uso en Diferentes Idiomas y Culturas: Fuentes de Error, Posibles Soluciones y Directrices Prácticas. En: Muñiz J. Psicometría. Madrid: Editorial Universitas, S. A; 1996. p. 207-238.
13. Prieto AJ. A Method for Translation of Instruments to other Languages. Adult Education Quarterly 1992; 43:1-14.
14. Streiner DL. A Checklist for Evaluating the Usefulness of Rating Scales. Can J Psychiatry 1993; 38:140-148.
15. Feinstein A. Clinimetrics. Yale University Press; 1987.
16. Nunally JC. Psychometric theory. New York: McGraw Hill: 1978.
17. Norman GR, Streiner DL. Componentes Principales y Análisis de Factores. En: Norman GR, Streiner DL. Bioestadística. Madrid : Mosby-Doyma Libros; 1996. p. 129-142.
18. Campbell DT, Fiske AW. Convergent and Discriminant Valdation by Multitrait-multimethod matrix. Psychological Bulletin 1959;56:81-105.

19. Streiner D, Norman GR. Health Measurement Scales. A Practical Guide to their Development and Use. 2nd ed. Oxford: Oxford University Press; 1995.
20. Guyatt HG, Walter S, Norman G. Measuring Change Over Time: Assessing the usefulness of Evaluative Instruments. J Chronic Dis 1987; 40:171-178.
21. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297-334.
22. Lin L. I-K.. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989; 45: 255-68.
23. Donner A, Eliasziw M. Sample Size Requirements for Reliability Studies. Statistics in Medicine 1987; 441-448.