Instrumentos de evaluación subjetiva en salud mental

L. SALVADOR* y M. ROCA**

* Profesor titular de Psiquiatría y Psicología Médica de la Universidad de Cádiz. ** Profesor titular de Psiquiatría de la Universitat de les ille Ballears.

1. INSTRUMENTOS DE EVALUACION: CONCEPTOS BASICOS Y CLASIFICACION

La evaluación puede definirse como aquel proceso consistente en aplicar un método sistematizado para describir fenómenos u objetos. Su grado de sistematización puede ser muy variable, yendo desde la mera asignación de códigos preestablecidos hasta los sistemas de cuantificación mediante algoritmos. Aunque algunos autores diferencian entre medición (proceso de recogida de la información) y evaluación (interpretación de los resultados), se tiende a denominar como «evaluación» todo el conjunto del proceso. La evaluación puede ser subjetiva u objetiva. La evaluación subjetiva se caracteriza por la descripción de constructos hipotéticos o intangibles (ej.: ansiedad, depresión) en oposición a las entidades tangibles descritas por las ciencias experimentales como el peso o la altura (evaluación objetiva). En ciencias de la salud esta diferenciación no siempre es diáfana, ya que existe una enorme carga individual en la interpretación de pruebas complementarias complejas (histología, diagnóstico por la imagen, neurofisiología). Ello determina que muchas normas de calidad sean comunes a los instrumentos objetivos y subjetivos). La evaluación subjetiva es menos precisa, y ha sido infravalorada hasta muy recientemente, pero la creciente demanda de parámetros intangibles como la calidad de vida, la satisfacción, el apoyo, la autonomía o el nivel de discapacidad del sujeto, ha determinado que en la actualidad la utilización de estos instrumentos sea imprescindible en cualquier área de salud.

La evaluación puede ser descriptiva o cuantitativa. La evaluación cuantitativa consiste en la elaboración de reglas para asignar números a un fenómeno dado, con el fin de cuantificar uno o varios atributos del mismo. Las reglas son una serie codificada de procedimientos para la asignación de números. Al evaluar un fenómeno concreto, es importante situarlo dentro de un modelo categorial o dimensional, y en este segundo caso delimitar su carácter uni o multidimensional. Cuando un fenómeno complejo se considera en un marco multidimensional, deben delimitarse las dimensiones básicas sobre las que centrar la evaluación, ya que su número real es prácticamente inabarcable. Así, la calidad de vida puede considerarse como una categoría dicotómica (ausente/presente), o como un fenómeno multidimensional. En el caso de la calidad de vida y salud podemos considerar una serie de dimensiones básicas: funcionamiento general (psíquico/grado de bienestar, físico/grado de autonomía y funcionamiento sociolaboral), síntomas asociados al trastorno y a su tratamiento, síntomas de distrés psíquico y discapacidad. De forma adicional se puede añadir: dolor, funcionamiento sexual, relaciones con el personal sanitario, y así sucesivamente.

Los instrumentos de evaluación constan de un número variable de items. El ítem es la unidad básica de información de un instrumento de evaluación, y suele componerse de una pregunta y de una respuesta que generalmente es cerrada y permite una asignación de un código. Los instrumentos de evaluación pueden dividirse en una serie de amplios grupos de acuerdo con su complejidad.

En el primer grupo pueden situarse los *cuestionarios descriptivos* (ej.: cuestionarios sociodemográficos) y los *inventarios de síntomas* (inventario de efectos secundarios). Estos instrumentos no permiten una cuantificación de sus items y pueden considerarse como meras listas de comprobación o chequeo. En un segundo nivel (Bech y cols.) se encuentran las *escalas de evaluación.* Como su nombre indica, éstas permiten una escalación acumulativa de sus items, dando puntajes globales al final de la evaluación. Se componen de items individuales, cada uno de los cuales describe una característica bien definida del fenómeno evaluado. Su carácter acumulativo las diferencia de los cuestionarios de recogida de datos y de los meros inventarios de síntomas. En un tercer nivel se sitúan las *entrevistas estandarizadas.* Estas se clasifican en función de su objetivo (generales o específicas) y según el nivel de capacitación requerido para su administración, que a su vez depende de la estructuración en la formulación de las preguntas y la codificación de las respuestas (a mayor estructuración menor nivel de capacitación requerido para la administración). Las entrevistas estandarizadas pueden acompañarse de un sistema informatizado de corrección que permite la asignación de ciertos diagnósticos. Los *sistemas de diagnóstico estandarizado* constituyen el cuarto nivel. Estos proporcionan

una codificación de entidades nosológicas, con una descripción detallada de cada una de ellas a través de un glosario para facilitar el diagnóstico. Los sistemas de diagnóstico se denominan operativos cuando proporcionan una serie de reglas para el diagnóstico basadas en criterios de inclusión (presencia de un número mínimo de características del fenómeno para su diagnóstico), y de exclusión (despistaje de otras características no relacionadas con el fenómeno). Cuando los criterios de exclusión se refieren a la presencia de otras entidades sindrómicas se considera que el sistema es jerárquico, puesto que efectúa una jerarquía de las entidades nosológicas recogidas en el sistema para su diagnóstico diferencial. Si además permite la codificación de varias entidades o aspectos relacionados en diversos ejes, se considera que el sistema es multiaxial. Existen dos sistemas de diagnóstico operativo jerárquico y multiaxial en vigencia en el momento actual: el sistema de investigación de la CIE-10 y el DSM-IIIR. Para algunos autores los sistemas de diagnóstico no deben ser considerados como un instrumento de evaluación. Sin embargo, en su construcción y su utilización, los sistemas diagnósticos se ajustan a las reglas generales de la evaluación subjetiva estandarizada. En un sexto nivel podemos situar una serie de instrumentos de reciente diseño, que generalmente se fundamentan en una entrevista estandarizada. A diferencia de éstas, las *baterías compuestas* de evaluación constan de un número complejo de instrumentos diferentes: cuestionario de recogida de datos, escalas de evaluación incorporadas a la batería, entrevista estandarizada para recogida de síntomas pasados y/o del estado actual, y sistema informático para diagnóstico múltiple, que permite la codificación diagnóstica según sistemas diferentes. Existen en la actualidad dos baterías compuestas que se ajustan a la descripción anterior: la batería SCAN, desarrollada a partir del PSE (Pull y Wittchen, 1991) (Vázquez-Barquero, 1993), y la batería CASH para evaluación de esquizofrenia y trastornos del estado de ánimo, desarrollada a partir del SANS/SAPS para evaluación de síntomas positivos y negativos en la esquizofrenia, entre otros instrumentos (Andreasen et al., 1992).

2. BASES PARA EL ESTUDIO DE LAS ESCALAS DE EVALUACION

Las escalas de evaluación son los instrumentos más numerosos y dispares utilizados en evaluación subjetiva. Para comprender las bases y el desarrollo de estos instrumentos nos basamos, por lo tanto, en este subgrupo. A continuación se comentan una serie de aspectos relacionados con sus fundamentos, el diseño y los parámetros utilizados en su construcción y en la evaluación su calidad. La utilización de escalas de evaluación se basa tanto en la psicofísica como en la psicometría. La psicofísica nos aproxima al proceso de la cuantificación de la percepción. Para trasladar a un sistema numérico fenómenos intangibles, como los síntomas psicológicos o la discapacidad, debemos establecer analogías. Dado que conocemos los exponentes en términos de juicios numéricos para diferentes estímulos (longitud linear, intensidad del sonido, brillo, presión, etc.), es posible establecer, a partir de juicios subjetivos, analogías escalares entre diferentes percepciones. La psicofísica da así una base para la utilización de determinadas escalas, como los análogos lineares de dolor (McDowel y Newell, 1987).

La psicometría nos permite conocer el grado en que una escala determinada se ajusta a un modelo matemático (adecuación matemática), y la forma en que se ordenan sus componentes o items (escalación numérica).

a) Jerarquía de adecuación matemática de una medida

1) Escalas nominale

En éstas, la asignación de los números es arbitraria y no tiene ninguna implicación de orden inherente (ej.: números de teléfono, estado civil). Los números se utilizan para clasificación, y el único análisis estadístico que permiten es el análisis canónico cuando se trata de items categoriales binarios. Su expresión matemática es:

$$A = B \text{ ó } A \neq B$$

2) Escalas ordinale

Los números se asignan como marcas para diferentes categorías, pero en este caso reflejan un orden creciente de la característica evaluada (ej.: los números de una calle, la depresión leve, moderada y marcada). Sin embargo, el valor real de los números y la distancia numérica entre cada categoría (Bech y cols.) no guardan ninguna relación intrínseca (ej.: un cambio de 2 a 3 no es equivalente al cambio de 3 a 4). El análisis estadístico que permite estas escalas no es paramétrico, ya que prácticamente nunca se cumplen las condiciones estrictas para utilizar estadística paramétrica. La suma global de las puntuaciones de una escala ordinal y el uso matemático de este valor numérico (ej.: puntuación global en el test de Hamilton o de Goldberg) es objeto de un intenso debate entre los metodólogos. Y debe tenerse en cuenta que puede llevar a conclusiones erróneas. La inmensa mayoría de escalas de evaluación psiquiátrica son escalas ordinales. Su expresión matemática es:

$$A > B \text{ ó } B > A$$

3) Escalas de intervalo

Los números siguen un intervalo constante a través de toda la escala. Así, la distancia entre una unidad y la siguiente en la parte inferior de la escala es igual a la distancia entre cada unidad en otras regiones de la escala

(ej.: Escala de Celsius para la temperatura). Permite la adición y resta de números pero no la multiplicación o el cociente. Permite el uso de estadística paramétrica. Su expresión matemática es:

$$A - B = C - D$$

4) Escalas de cociente

Escala de intervalo con un valor «0» real, lo que permite definir cocientes (saber cuántas veces mayor es una puntuación que otra) (ej.: peso, altura, salario). Permite cualquier tratamiento estadístico. Su expresión matemática es:

$$A \times B = C \quad y \quad C/B = A$$

b) Jerarquía de ordenación de items. Métodos de escalación numérica

Existen diversos métodos para convertir los indicadores descriptivos de una respuesta en estimaciones numéricas. El orden jerárquico de una serie de categorías dentro de un ítem o de un grupo de items en una escala ordinal, se puede establecer a través de un análisis de escalograma de Guttman. Este método ordena los items de forma que una respuesta positiva en uno de ellos implica una respuesta positiva en todos los que le preceden (ej.: «Puedo caminar un kilómetro», «Puedo caminar una manzana», «Puedo caminar fuera de mi casa», «Puedo moverme en mi habitación»). El análisis de escalograma según el modelo de Rasch permite obtener un análisis más preciso de la estructura latente del test (Bech y cols., 1993).

Las categorías de cada ítem en una escala ordinal suelen oscilar entre 2 y 6 (ausente, dudoso, leve, moderado, marcado y severo). La puntuación depende del patrón de asignación numérica que se seleccione, y que se detalla en el apartado 2.2.2. (composición de las escalas de evaluación).

3. ESCALAS DE EVALUACION

3.1. Clasificación

Bech y cols. (1993) proponen una descripción basada en los objetivos y la composición de la escala:

1) *Area de evaluación:* escalas diagnósticas, sintomáticas, de personalidad y escalas para otros propósitos específicos.

2) *Tipo de administración:* escalas para el paciente, el médico u otro personal sanitario.

3) *Acceso temporal retrospectivo:* marco temporal de la evaluación.

4) *Selección de items:* distingue entre escalas de primera generación (basadas en experiencia clínica), y de segunda generación (derivadas de las anteriores).

5) *Número de items de la escala,* y

6) *Definición de los items individuales.*

En base a las descripciones efectuadas por otros autores (cfr. Thompson, 1989; Wittchen y Essau, 1991), se ha modificado la propuesta original de Bech para permitir una descripción más completa de las diferentes escalas. Los cambios en la terminología con respecto a la utilizada por Bech se detallan en cada apartado.

Al diseñar una escala de evaluación es fundamental tener en cuenta el propósito de la misma en sus diferentes aspectos (patología evaluada, población de referencia, periodo de evaluación, etc.), la composición de sus items y la prevención de sesgos potenciales en su cumplimentación. Se detallan a continuación estos factores, aportando algunos ejemplos (las siglas corresponden a escalas mencionadas en la tabla IV).

3.1.1. Propósito de la escala

El propósito de una escala va a determinar el contenido de sus items y diversos aspectos relacionados con su estructura. Una escala debe limitarse siempre al área para la que ha sido diseñada, a menos que se efectúe una nueva estandarización de la misma. El propósito se relaciona con la dimensión evaluada, la población objeto de estudio, el periodo de evaluación y el tipo de cumplimentación.

a) Area evaluada

Las diversas escalas psicosociales evalúan un amplio rango de áreas como síntomas (escalas clínicas), personalidad, adaptación social, familiar, sexual, laboral, discapacidad, etc. Bech (1993) efectúa una distinción entre dos tipos de escalas clínicas: de diagnóstico y sintomáticas. Esta distinción es conflictiva, al existir escalas sintomáticas que han sido utilizadas para diagnóstico tras calcular el punto de corte idóneo a través de un estudio de validez predictiva (ver parámetros de calidad de una escala), y viceversa.

b) Objetivo de estudio

Este nos permite diferenciar entre escalas generales (p. ej. para evaluación de caso psiquiátrico) y escalas específicas (p. ej. para evaluación de depresión). Las escalas específicas pueden tener a su vez diferentes gradaciones (ej.: para evaluación de depresión mayor, y escala de Newcastle para evaluación de depresión endógena). Wittchen y Essau (1991) distinguen entre escalas basadas en un concepto «amplio» o «restrictivo» de trastorno mental. Los instrumentos más restrictivos priman la especificidad sobre la sensibilidad y viceversa

(este factor es particularmente importante en el uso de sistemas de diagnóstico estandarizado).

Marc temporal

En función de la *estabilidad* del fenómeno evaluado podemos diferenciar entre las escalas de rasgo, que evalúan fenómenos relativamente estables a lo largo del tiempo (p. ej.: test de personalidad, locus de control); y escalas de estado, que evalúan la situación actual del sujeto –generalmente en el último mes– (p. ej. depresión, síntomas negativos y positivos), las últimas semanas o la semana anterior, o los tres días anteriores a la evaluación (escalas de «aquí y ahora»). El marco temporal debe detallarse en las instrucciones previas a la administración de la escala.

En las escalas de estado, el *período de evaluación* nos permite diferenciar entre escalas de detección (p. ej. para identificación del caso psiquiátrico –GHQ–), escalas de seguimiento no transicionales y transicionales. Las escalas de seguimiento no transicionales evalúan el cambio en función de la diferencia del puntaje entre dos evaluaciones (ej.: HDS), mientras que las transicionales evalúan directamente el grado de mejora o empeoramiento experimentado por el paciente entre ambas evaluaciones (ej.: escala de cambio de CGI). En el estudio de una escala de seguimiento es importante conocer su sensibilidad al cambio.

d) Tipo de administración

Las escalas autoadministradas se diseñan para ser cumplimentadas por el propio sujeto o por un informante. En ocasiones se incluyen items para calibrar la validez de las respuestas en función de la tendencia a disimulación o a simulación (ej.: EPQ de Eysenck). Bech y cols. (1993) denomina a este grupo de instrumentos «cuestionarios», sin embargo este término es demasiado amplio.

Las escalas heteroadministradas («escalas de observador» según Bech), son cumplimentadas por el evaluador del estudio. Los instrumentos de evaluación heteroadministrados requieren diferentes niveles de capacitación profesional para su uso (este factor es particularmente importante en el diseño y administración de entrevistas estructuradas). Las escalas heteroadministradas requieren una estandarización previa del examinador a través de un análisis de su acuerdo con un examinador de referencia (ver fiabilidad interexaminadores). Se señalan dos tipos de situaciones extremas en la administración de una escala heteroadministrada: *situación Alfa* (investigador experto que sigue un interrogatorio cerrado y utiliza una escala con pocos items, bien definidos, y que incluyen criterios de mejoría y de salud); y *situación Beta* (evaluador inexperto, que realiza una entrevista abierta, y utiliza una escala con muchos items mal definidos y sin criterios de mejoría y salud) (cfr. Bech y cols., 1993).

Algunos instrumentos de evaluación clínica son de tipo mixto, incluyendo una sección para síntomas referidos y otra distinta para síntom observados en la ent vista.

onstrucción de as alas de e

Como ya se ha in
de información de u ıment
neralmente consta d ɔregun ɔpue
cerrada.

a) Número de items

Puede distinguirse entre escalas unitarias o globales, compuestas de un solo ítem (ej.: CGI, GAS, HDS, escalas analógicas de dolor o de bienestar); y escalas multi-ítem. Como regla general, se considera que un fenómeno debe ser evaluado con un mínimo de 6 items (Bech y cols., 1993). Generalmente las escalas constan de entre 10 y 90 items. Diversas escalas están disponibles en varias versiones. Así, el GHQ de Goldberg puede utilizarse en su versión de 60, de 30, de 28 o de 12 items; y el HDS de Hamilton en versiones de 21 o 17 items (aparte de otras escalas derivadas de esta prueba).

ɔntenid de los iten

En función del contenido se distingue entre escalas unidimensionales y multidimensionales. En las escalas unidimensionales, más del 80% de los items evalúan una sola dimensión de acuerdo con el modelo de Israel (1983): *Dimensión física* (síntomas relacionados con aspectos médicos, corporales –ej.: cuestionario de dolor de McGill), *dimensión psíquica* (aspectos cognitivos –BDI–), y *dimensión social* (ej.: SAS, ADL). En las escalas multidimensionales los items evalúan dos o tres de las dimensiones señaladas (ej.: GHQ, HDS).

El sesgo de ítem u orientación se refiere a la parte del síndrome que aparece mejor reflejada en la escala, y se representa en un porcentaje de la puntuación máxima teórica para cada categoría de síntomas (Thompson, 1989).

) Definición y orden de s item

La definición de cada ítem debe ser exhaustiva y mutuamente excluyente (criterios de Guilford) (cfr. Bech y cols., 1993). Por otro lado, deben tenerse en cuenta una serie de factores tanto al formular las preguntas y las alternativas de respuesta, como al ordenar el conjunto de items que componen la escala:

1) *Comprensión*. Es necesario adaptar el lenguaje y el tipo de formulación de las preguntas y respuestas al entorno sociocultural del paciente. Así, por ejemplo, la comprensión del uso de análogos lineares tiende a ser mejor en el medio anglosajón que en la Europa meridional, donde la comprensión de análogos numéricos decimales es mayor. Existen diversos índices de evaluación

de la comprensibilidad de un texto (ej.: índice de Flesch para el idioma inglés). El problema de la comprensión es sumamente importante en la evaluación de poblaciones específicas como la de los sujetos afectos de retraso mental. Por otro lado, la traducción y adaptación de una escala previamente desarrollada en otro idioma y entorno cultural debe seguir una tecnología específica que incluya un proceso de retrotraducción. Recientemente se han aplicado sistemas más complejos como el de traducción conceptual.

2) *Aceptabilidad.* Es fundamental que los items sean aceptables para el sujeto evaluado. La desideratividad social es un tipo de sesgo potencial que puede alterar la validez de los resultados en las respuestas (Wittchen y Essau, 1991), y que debe tenerse en cuenta al formular las preguntas de determinados items (este sesgo es importante en la evaluación de las actitudes ante determinadas enfermedades como el SIDA, en las que el sujeto tiende a responder aquello que considera como socialmente más aceptable). También es necesario limitar el número de items para evitar el cansancio y favorecer la colaboración del sujeto (este problema es evidente en cuestionarios o en baterías de más de 100 items como el MMPI).

3) *Prevención de sesgos en la cumplimentación.* La aquiesencia (tendencia a responder afirmativamente a la pregunta) determina la necesidad de alternar preguntas formuladas «en negativo». Sin embargo, este tipo de formulación puede disminuir significativamente la comprensión del paciente y la fiabilidad de las respuestas (p. ej.: items del tipo: «No es cierto que Colón descubrió América» = V/F). El error de tendencia central se refiere a la reticencia a responder las alternativas extremas en un ítem, dando preferencia a las centrales. Este problema afecta (Bech y cols.) principalmente a las escalas analógico-verbales de tres o cinco alternativas (ej.: nada, algo, mucho). Otro tipo de sesgo se relaciona con la tendencia a responder más a las alternativas situadas a la derecha o a la izquierda, lo que se incrementa cuando uno de los dos extremos contiene siempre las alternativas «deseables», y puede evitarse alternando primero items con alternativas positivas a la izquierda y después items con alternativas positivas a la derecha.

Cuando se diseña una escala heteroadministrada (cumplimentada por el evaluador), deben tenerse en cuenta algunos sesgos específicos: El efecto halo se refiere a la tendencia a efectuar un juicio al inicio de la entrevista (ej.: diagnóstico heurístico) que condiciona la cumplimentación de los items siguientes (ello puede acontecer en el HDS, que agrupa los items directamente relacionados con depresión y severidad al inicio de la entrevista). El error lógico se produce al juzgar que todos los items aparentemente relacionados deben puntuarse de forma similar (así, puede asumirse que un paciente con una puntuación elevada en «ideas suicidas» puntuará también alto en «desesperanza»). El error de proximidad conduce a puntuar de forma similar los items adyacentes. Otra fuente de error es la varianza terminológica que se relaciona con la atribución de un significado diferente a un mismo término. Este problema afecta sobre todo a las escalas clínicas, dada la diferente interpretación de un término según la escuela psicopatológica o los conocimientos de base del evaluador. Este sesgo puede obviarse incluyendo un glosario terminológico anexo a la escala de evaluación (ej.: BPRS).

d) Selección de los items

Heelh y Golden (1982) señalan una serie de principios o pasos en la construcción de una escala de evaluación de síntomas:

1. Selección de los items en función de su relevancia clínica y validez.
2. Seleccionar los items en función de la correlación interna de los items cuando se aplican a un grupo mixto de pacientes (que incluye a pacientes con y sin el síntoma evaluado).
3. Selección de items con diferente peso jerárquico (que describan los diversos aspectos del fenómeno evaluado).
4. Ante igualdad de factores, seleccionar los items con mayor potencial de consenso.
5. Comprobar el rendimiento del grupo de items seleccionado en función de diversos criterios externos (edad, sexo, etc.), con el fin de evaluar su transferibilidad.
6. Cuando los pasos 3, 4 y 5 no se puedan efectuar, repetir el análisis con items modificados HDS en cuanto a definición o contenido.

Además de la selección según el propósito y la selección matemática, los items pueden seleccionarse en función de su utilidad. Esta se evalúa de acuerdo con tres criterios (Thompson, 1989):

1. *Calibrado:* frecuencia suficiente de respuestas en un ítem individual como para garantizar su inclusión en la escala. Arbitrariamente puede fijarse en un 10%.
2. *Monotonicidad ascendente:* El ítem debe mostrar una correlación significativa con la puntuación global (ver homogeneidad).
3. *Baja dispersión* con respecto a la línea de regresión de la correlación anterior.

e) Sistema de codificación de respuesta

1) *Escalas categoriales dicotómicas.* Presentan un sistema de respuesta de dos alternativas: Sí/No o Verdadero/Falso (ej.; test de personalidad como el EPQ o el MMPI).

2) *Escalas analógicas.* Pueden diferenciarse en función del sistema analógico utilizado para facilitar la respuesta:

Escala analógico-linear: Gradación en una línea de 7 a 10 cm (ej.: escalas para dolor o bienestar).

Ejemplo:

Ningún dolor ——————————— El máximo dolor imaginable

Escala analógico-numérica: Gradación similar a la anterior pero con números (de 0 a 7 o a 10). En las escalas unitarias termométricas los números se colocan en posición vertical. Estas pueden también graduarse de 0 a 100 (ej.: GAF para evaluación de funcionamiento general psíquico). En ocasiones se combinan análogos visuales y numéricos para aumentar la comprensión (ej.: escala de calidad de vida de la EORTC, 1986).

Ejemplo:

Ninguna ansiedad **0 1 2 3 4 5 6 7** Máxima ansiedad imaginable

Escalas gráficas. Gradación a través de dibujos (ej.: *face scale* para evaluación de bienestar). Algunos autores consideran a las escalas gráficas como escalas lineares.

Escala analógico-verbal. Gradación en categorías verbales previamente calibradas (ver escalación de Guttman). Generalmente las opciones de respuesta oscilan entre 3 y 7. Likert consideraba que 5 era el número de alternativas óptimo. Goldberg por su parte, prefiere utilizar cuatro grados de respuesta para evitar el sesgo de tendencia central (ver apartado 3.2.3.a3). Se considera que por encima de 6 grados, el nivel de fiabilidad disminuye significativamente. Las escalas de severidad usan más grados que las de detección (ej.: el CGI tiene 7 mientras que el GHQ tiene 4). Estas escalas también reciben el nombre de *Likert* en honor a su introductor, hace 60 años (Cfr. Bech y cols., 1986). Sin embargo, también se denomina así un sistema específico de puntuación, por lo que este uso puede prestarse a confusión.

Ejemplo:

No más de lo habitual	*Algo menos de lo habitual*	*Bastante más de lo habitual*	*Mucho más de lo habitual*

E. analógico-categoriales: Se consideran dentro de este grupo una serie de escalas que combinan gradación numérica y verbal (ej.: CGI, GAF) (Bech y cols., 1986). También se conocen como escalas DISCAN (*Discretized Analogue Scale*).

f) Puntuación de los items

El sistema de puntuación puede variar sustancialmente de una escala a otra, e incluso en una misma escala, cuando se trata de escalas analógico-verbales.

Las escalas unitarias de severidad (no transicionales), suelen tener una puntuación máxima de 8 o 10 cuando son análogos visuales, y de 7 a 10 cuando se trata de análogos verbales u otras formas combinadas (DISCAN). El GAF puede puntuarse hasta 99, pero en realidad presenta 10 grados de respuesta en decimales.

Las escalas globales unitarias de tipo transicional generalmente son de tipo bipolar, permitiendo una puntuación en sentido negativo y positivo (de mayor empeoramiento a mayor mejoría). Por razones técnicas también pueden puntuarse de 1 a 7, aunque la polaridad de la escala no queda adecuadamente reflejada en este sistema.

- *Alternativa A:* –3 / –2 / –1 / 0 / 1 / 2 / 3
- *Alternativa B:* 1 / **2** / 3 / 4 / 5 / 6 / 7

Las escalas verbales multi-ítem permiten diversas asignaciones numéricas. Así el GHQ de Goldberg permite tres asignaciones diferentes: las dos primeras en base al sistema originariamente propuesto por Likert en los años treinta, y un tercero propuesto por el propio Goldberg. El HDS y el SANS/SAPS se puntúan de acuerdo con el sistema de puntuación propuesto por M. Hamilton, que disingue la opción de ausente (0), dudoso (1), y diversos grados de intensidad (del 2 al 4 o 5).

– Goldberg	0 - 0 - 1 - 1
– Likert I	0 - 1 - 2 - 3
– Likert II	0 - 0 - 1 - 2
– Hamilton	0 - 1 - 2 - 3 - 4

3.2. Parámetros de la calidad de un instrumento de evaluación

Existen tres parámetros básicos para evaluar la calidad de un instrumento de medida: su consistencia, su fiabilidad y su validez.

3.2.1. Consistencia

La consistencia analiza el nivel en que los diferentes items de una escala están relacionados entre sí. Algunos autores incluyen a la consistencia dentro de la fiabilidad o de la validez.

a) *Consistencia interna:* La homogeneidad indica el grado de acuerdo entre los items de la escala, lo que determina si éstos pueden acumularse y dar una puntuación global. Esta se puede obtener mediante el estudio de la correlación de los items con el total (partición media, coeficiente alfa de Cronbach), análisis factorial o mediante los modelos de objetividad estadística de Rasch (1980). La *partición media* (split-half) estima la homogeneidad en función de la correlación entre dos mitades equivalentes de la escala (ej.: items de primera mitad versus items de segunda mitad, o items impares versus items pares). El coeficiente alfa de Cronbach indica el grado en que los diferentes items presentan una correlación positiva (la consistencia interna es alta por encima de 0,7) (Bech y cols., 1993). La homogeneidad a partir del análisis factorial (aceptabilidad de la puntuación global como suma de la obtenida en cada ítem), se confirma si se obtiene una estructura unidimensional, esto es, todos los items muestran una carga positiva en el primer factor (Thompson, 1989). El modelo de Rasch unidimensional

considera que una escala es homogénea cuando todos sus ítems contribuyen de forma independiente al total de información contenido en la escala.

b) *Consistencia jerárquica:* Esta viene determinada por una serie de factores: el coeficiente de escalabilidad se refiere al grado en que la escala es acumulativa. El modelo de Rasch permite también estudiar la jerarquía interna de la escala, clasificando los ítems homogéneos en un rango jerárquico del más inclusivo (que mide los síntomas leves o moderados de la dimensión) al más exclusivo (que mide los síntomas más graves de la dimensión). El coeficiente de reproducibilidad indica hasta qué punto la escala refleja todos los patrones de respuesta del sujeto con respecto al parámetro medido (Thompson, 1989). La transferibilidad se refiere al grado en que la escala puede ser aplicada a diferentes grupos de población que presenten el fenómeno evaluado, independientemente de la edad, sexo y otros criterios externos relevantes (Bech y cols., 1993).

3.2.2. Fiabilidad

La fiabilidad nos indica el grado en que los resultados de un test son reproducibles. Esta medida depende de la estabilidad de las medidas del test a pesar del cambio de diversos parámetros externos. El estudio de la fiabilidad externa informará sobre la reproducibilidad de los resultados del test en distintas situaciones. McDowell (1987) ejemplifica la diferencia entre la validez y la fiabilidad con un excelente símil: un tirador tiene que aprender a acertar en el blanco y después hacerlo de forma consistente. La validez vendría dada por el grado en que el disparo se acerca a la diana, y la fiabilidad por el grado de aproximación entre sí de una serie consecutiva de disparos.

Un estudio sobre la fiabilidad de una prueba diagnóstica debe incluir al menos un análisis del nivel de acuerdo obtenido al ser evaluada la misma muestra en las mismas condiciones por dos evaluadores distintos (fiabilidad interexaminadores), y la concordancia obtenida entre los resultados del test al ser evaluada la misma muestra por el mismo evaluador en dos situaciones distintas (fiabilidad test-retest). En algunos casos, como en psiquiatría infantil, minusvalías psíquicas, etc., se recaban (Bech y cols.) los datos obtenidos con el test con la misma muestra y con el mismo evaluador, pero recabando los datos de dos informadores distintos (fiab. inter-informadores).

El índice estadístico utilizado para evaluar la concordancia depende de las características de las variables a evaluar. En el caso de variables dicotómicas o binarias, la concordancia ítem a ítem se puede analizar mediante el porcentaje de acuerdo y el kappa no ponderado (Kramer y Feinstein, 1981). El coeficiente de concordancia kappa nos informa del nivel de acuerdo obtenido, una vez eliminada la concordancia que presumiblemente se ha producido por azar. Ello lo hace más fiable que el simple porcentaje de acuerdo. Feinstein (1985) propone la siguiente tabla para analizar los resultados del kappa:

Valor de kappa	*Nivel de acuerdo*
<0	Pobre
0 - ,20	Bajo
,21 - ,40	Regular
,41 - ,60	Moderado
,61 - ,80	Fuerte
,81 - 1,00	Casi perfecto

En el caso de variables ordinales, el análisis de la concordancia ítem a ítem se puede efectuar utilizando el porcentaje de acuerdo ponderado y el kappa ponderado. Estos se consideran más adecuados que sus análogos no ponderados por dar una medida más real del nivel de acuerdo al ponderar el desacuerdo según el número de rangos que separaran la puntuación asignada por un evaluador de la asignada por el otro (así, el peso asignado puede ser 0 para el completo acuerdo, 1 cuando hay un rango de diferencia, 2 cuando hay dos rangos, etc.) (Kramer y Feinstein, 1981).

El método de análisis de la concordancia de las puntuaciones globales de un test es controvertido. Habitualmente se utilizan los coeficientes de correlación para analizar el grado de acuerdo. Dichos coeficientes no deberían ser utilizados para analizar la concordancia entre dos evaluaciones: la tendencia puede ser perfecta, con un coeficiente de correlación de 1, y las medidas obtenidas en un laboratorio ser exactamente el doble o exactamente 10 unidades más altas que las correspondientes medidas obtenidas en otro laboratorio, con lo que la concordancia obtenida entre ambos laboratorios sería inexistente (Feinstein, 1985). En medidas continuas se puede utilizar el coeficiente de correlación intraclase (ICC). Bech y cols. (1993) aconsejan también la utilización del IC para la evaluación de la fiabilidad test-retest cuando las medidas han sido efectuadas por evaluadores diferentes. En actualidad, no existe un acuerdo sobre el tamaño de la muestra requerido para un estudio de fiabilidad de una escala (Bech y cols., 1993).

3.2.3. Validez

La validez se define por el grado en que el instrumento mide aquello que realmente pretende medir. Existen múltiples formas de validez, con el agravante de que algunos autores utilizan un mismo término para definir conceptos diferentes. En la tabla I se recogen una serie de parámetros relacionados con la validez de los instrumentos de evaluación. Entre todos ellos, el más importante desde el punto de vista práctico es la validez predictiva de observación, que debe distinguirse de otros tipos de validez predictiva, como la de respuesta al tratamiento o la de desenlace.

La *validez predictiva de observación* se refiere a la probabilidad de la escala de dar un juicio correcto sobre

el fenómeno observado. El análisis de Bayes permite conocer la validez predictiva de un test, su utilidad y su comparabilidad, a partir del análisis de la distribución de los «casos» y «no casos» en una población dada y de su relación con los resultados obtenidos en el test estudiado (positivos y negativos). Una tabla de contingencia de 2x2 expresa esta relación en Verdaderos Positivos (VP), Verdaderos Negativos (VN), Falsos Positivos (FP) y Falsos Negativos (FN). En la tabla II se definen los coeficientes de validez predictiva obtenidos a partir de la tabla de contingencia: sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo. Otros parámetros que pueden obtenerse a través de la aplicación del teorema de Bayes son la proporción de mal clasificados, la eficiencia (proporión de casos bien clasificados en relación al total de evaluados), el sesgo (cociente entre los evaluados considerados como positivos y negativos) y el rendimiento (casos no detectados por el test en relación al total de casos). Estos coeficientes nos permiten ajustar el punto de corte con respecto al objetivo del estudio. Si se pretende hacer un estudio de muestreo en dos fases. buscaremos aquel punto de corte que nos permita captar el máximo número de casos aunque entre ellos se incluyan falsos positivos (especificidad aceptable con una sensibilidad óptima). Si por el contrario pretendemos conocer la morbilidad probable en una población a través de la puntuación en el test, seleccionaremos aquel punto de corte que nos permita descartar el mayor número de «no casos», aunque ello determine la pérdida de algunos falsos negativos (sensibilidad aceptable con una especificidad óptima).

De forma alternativa, el punto de corte idóneo de un test se puede calcular a partir del análisis ROC (*Receiver Operating Characteristics*). Esta técnica fue desarrollada en los años sesenta para evaluar la capacidad de discriminación de señales de los controladores de radar. Primero se obtiene una representación gráfica de la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos (1-especificidad) para cada punto de corte. El cálculo del área bajo la curva resultante nos indica la capacidad discriminante del test a través de todo el *spectrum* de morbilidad. Cuando la capacidad discriminante es igual a la obtenida aleatoriamente se obtiene una línea diagonal cuya área inferior es de 0,5 (sensibilidad igual a la tasa de falsos positivos). Un test ideal produciría un 100% de verdaderos positivos antes de admitir un solo falso negativo, por lo que el área bajo la curva obtenida sería de 1,0. En la práctica las áreas bajo la curva oscilan entre 0,5 y 1,0, y permiten una representación gráfica de la capacidad discriminante de diferentes test para una misma dimensión, siendo el mejor aquel que se corresponda con una curva más alejada de la diagonal (fig. 1) (cfr. Thompson, 1989).

La validez predictiva de un test autoadministrado puede verse reducida como resultado de un error en el criterio de referencia. Los modelos de análisis factorial permiten solventar este problema, mediante el cálculo de la validez factorial. Se asume que los diferentes instrumentos (test de detección, criterio de referencia) miden un mismo constructo que puede representarse por un factor único resultante del análisis factorial de máxima probabilidad de las puntuaciones obtenidas en los diferentes instrumentos. Así, cada medida pueden compararse con el factor, que se toma como criterio, tanto para el cálculo de la fiabilidad como para el análisis de la capacidad discriminante de las diferentes pruebas en el análisis ROC. El estudio de la homogeneidad y la transferibilidad de una escala a través del modelo de objetividad estadística de Rasch, es considerado como parámetro de validez por algunos autores (Thompson, 1989).

3.3. Escalas de evaluación en psiquiatría

Dado su número, diversidad y su continuo desarrollo, es cada vez más necesario contar con inventarios informatizados y sistemáticos de las escalas de evaluación psiquiátrica. Una guía para su clasificación puede obtenerse en función de siete parámetros relacionados con su diseño (objetivo, tipo de administración, estabilidad del fenómeno, periodo de evaluación, número de items, contenido de los items y sistemas de respuesta) (tabla III). En la tabla IV aparece la relación de una serie de escalas psiquiátricas en función de su objetivo y tipo de administración. En el listado de escalas específicas sólo se incluyen aquellas para evaluación de trastornos afectivos y esquizofrenia, dada la imposibilidad de hacer una revisión extensa de estos instrumentos. El uso de estos instrumentos está bien establecido en diversas áreas, desde la epidemiología clínica a los estudios farmacológicos (propósito para el que se desarrollaron muchas de las escalas clínicas de seguimiento).

BIBLIOGRAFIA

1. ANDREASEN, N. C.; FLAUM, M. y ARNDT, S.: «The comprehensive assessment of symptoms and history (CASH): An instrument for assessing diagnosis and psychopathology». Archives of General Psychiatry, 49: 615-623, 1992.
2. BECH, P.; KASTRUP, M. y RAFELSEN, O. J.: «Mini compendium of rating scales for states of anxiety, depression, mania, schizophrenia with corresponding DSM-III syndromes». Acta Psychiatr. Scand., 73 (326), 1986.
3. BECH, P.; MALT, U. F.; DENCKER, S. J.; AHLFORS, U. G.; ELGEN, K.; LEWANDER, T.; LUNDELL, A.; SIMPSON, G. M.; LINGJAERDE, O. (eds.): «Scales for assessment of diagnosis and severity of mental disorders». Acta Psychiatrica Scandinavica, 87 (372), 1993.
4. CONDE, V. y FRANCH, J. I.: «Escalas de evaluación comportamental para la cuantificación de la sintomatología psicopatológica en los trastornos angustiosos y depresivos». Upjohn Publ., Madrid, 1984.

5. DERROGATIS, L. R. y WISE, T. N.: «Anxiety and depressive disorders in the medical patient». American Psychiatric Press. Washington, 1989.
6. EORTC Study Group on Quality of Life. Protocol 15861. EORTC data center. Brussels, 1986.
7. FEINSTEN, A. R.: «Clinical epidemiology». WB Saunders. Philadelphia, 1985.
8. GUY, W.; ECDEU: «Assessment Manual for Psychopharmachology». U.S. Dept. Health Education and Welfare. NIMH, Maryland, 1983.
9. ISRAEL, L.; KOZAREVIC, D. y SARTORIUS, N.: «Source book for the geriatric assessment: I. Evaluation in gerontology». World Health Organization. Karger. Basel, 1984.
10. KARNOFSKY, D. A. y BURCHENAL, J. H.: «The clinical evaluation of Chemotherapeutic agents». En: Columbia University Press. Evaluation of chemotherapeutic agents. New York, 1949.
11. KRAMER, D. A. y FEINSTEIN, A. R.: «Clinical biostatistics: LIV. The biostatics of concordance». Clin. Pharmacol. Ther., 29: 111-123, 1981.
12. LIKERT, R.: «A technique for measurement of attitudes». Archives of Psychology, 140: 1-55, 1932.
13. McDOWELL, I.; NEWELL, C.: «Measuring Health: A guide to rating scales and questionnaires». Oxford University Press. Oxford, 1987.
14. MEEHL, P. y GOLDEN, R. R.: «Taxonometric methods». En: Kendall, P. C.; Butcher, J. N. (eds.): Handbook of research methodology in clinical psychology. Wiley & Sons. New York, 1982.
15. MEZZICH, J. E.: «Clinical Care and Information Systems in Psychiatry». American Psychiatric Press. Washington, 1986.
16. PULL, C. B. y WITTCHEN, H. U.: «The CIDI, SCAN, and IPDE: Structured diagnostic interviews for ICD-10 and DSM-III-R». European Psychiatry, 6: 227-285, 1991.
17. RASCH, G.: «Probabilistic models for some intelligence and attainment tests». University of Chicago Press. Chicago, 1980.
18. SALVADOR, L.: «Instrumentos de evaluación en Psiquiatría». En: Bernardo, M.; Cubí, R. (dirs.): Detección de trastornos psicopatológicos en atención primaria. Societat Catalana de Medicina Psicosomática. Barcelona, 1989.
19. SALVADOR, L.: «Escalas de evaluación en Oncología Psicosocial». Monografías de Psiquiatría, 2 (5): 7-11, 1990.
20. THOMPSON, C. (dir): «The instruments of Psychiatric research». John Wiley & Sons. Chiclester, 1989.
21. VAZQUEZ-BARQUERO, J. L. (dir.): «SCAN. Cuestionarios para la evaluación clínica en Psiquiatría». Meditor. Madrid, 1993.
22. WARE, J. E.: «Measuring health and functional status in mental health services research». En: Tabue, C. A.; Mechanic, D. y Hohmann, A. (eds.): Department of Health & Human Services. Washington, 1989.
23. WARE, J. E. Jr. y SHERBOURNE, C. D.: «The MOS 36-item short form health survey (SF-36): I. Conceptual framework and item selection». Medical Care, 30: 473-483, 1992.
24. WELTZER, S.: «Medición de las enfermedades mentales: Evaluación psicométrica para los clínicos». Ancora. Barcelona, 1991.
25. WITTCHEN, H. U. y ESSAU, C. A.: «Assesment of symptoms and psychosocial disabilities in primary care». En: Sartorius, N.; Goldberg, D.; de Girolamo, G.; Costa e Silva, J.; Lecrubier, Y. e Wittchen, U. (dirs.): Psychological disorders in general medical settings. Hogrefe & Huber Publ. Toronto, 1990.