Capítulo 22

Validación de cuestionarios

Antes de generalizar la aplicación de un cuestionario es necesario evaluar su fiabilidad y su validez, lo que supone tiempo y esfuerzo. Por ello, cuando sea posible, un investigador debería usar cuestionarios que ya hubieran mostrado su utilidad, fiabilidad y validez en otros estudios, lo que además le permitiría comparar los resultados.

Sin embargo, hay ocasiones en que es inevitable el uso de nuevos cuestionarios: cuando los existentes han mostrado resultados poco satisfactorios, cuando un cuestionario se ha mostrado útil pero en un medio distinto, o cuando no hay ninguno que sea adecuado para medir las variables de interés. En estas circunstancias, es preceptivo evaluar la utilidad del nuevo cuestionario a partir de dos criterios: su fiabilidad y su validez.

Los cuestionarios deben poseer una serie de características: ser sencillos, viables y aceptados *(feasibility),* fiables, válidos y bien adaptados culturalmente, útiles y sensibles a los cambios. Mientras que la fiabilidad y la validez son exigencias necesarias en todos los instrumentos, la importancia de otras características psicométricas dependerá del contexto; así, por ejemplo, la sensibilidad al cambio *(responsiveness)* será muy importante si el instrumento se aplica como medida de la respuesta en los ensayos clínicos, pero no lo será tanto en un estudio sobre opiniones o actitudes acerca de una enfermedad.

La validación de un cuestionario es un proceso complejo que implica múltiples fuentes de información y la recogida de diferentes evidencias empíricas. Dado que no existe un criterio concreto ni único a partir del cual pueda considerarse que un cuestionario es válido, en general es necesario realizar más de un estudio con esta finalidad. Las características consideradas en la validación de un cuestionario se detallan en el cuadro 22.1.

VIABILIDAD

Los mejores instrumentos son inservibles si su aplicación resulta compleja y costosa. Características como el tiempo empleado en la cumplimentación del cuestionario, la sencillez y la amenidad del formato, y el interés, la brevedad y la claridad de las preguntas, así como la facilidad de la puntuación, el registro y la codificación, y de la interpretación de los resultados, son algunos aspectos relacionados con la viabilidad.

FIABILIDAD

Un instrumento es fiable si produce resultados consistentes cuando se aplica en diferentes ocasiones (estabilidad o reproducibilidad). Esquemáticamente, se evalúa administrando el cuestionario a los mismos sujetos, ya sea en dos ocasiones distintas (repetibilidad) o

Cuadro 22.1 Características que deben considerarse en la validación de un cuestionario

- Viabilidad.
- Fiabilidad:
  - Repetibilidad.
  - Fiabilidad interobservador.
  - Consistencia interna.
- Sensibilidad:
  - Sensibilidad al cambio.
- Validez:
  - Validez lógica.
  - Validez de contenido.
  - Validez de criterio.
  - Validez de constructo o de concepto.

por dos observadores diferentes (fiabilidad interobservador). Se trata, por tanto, de analizar la concordancia entre los resultados obtenidos en las diversas aplicaciones del cuestionario. Si la escala de medida es cualitativa, se evalúa mediante el índice kappa de Cohen, y si es cuantitativa, principalmente mediante el coeficiente de correlación intraclase (anexo 4). Otro concepto relacionado con la fiabilidad es el de la consistencia interna, que mide el grado en que se obtienen respuestas homogéneas a diferentes preguntas sobre un mismo concepto o dimensión. Siempre que sea posible deben evaluarse todos los componentes de la fiabilidad.

Repetibilidad

La repetibilidad, o fiabilidad test-retest, se refiere a si, cuando se administra un cuestionario a los mismos sujetos en dos ocasiones diferentes en el tiempo, se obtienen resultados idénticos o similares.

Las limitaciones de la evaluación de la repetibilidad de un cuestionario son las mismas que las de cualquier otro instrumento de medida. Si el tiempo transcurrido entre ambas aplicaciones del cuestionario es muy largo, el fenómeno que se mide puede haber presentado variaciones, mientras que si es demasiado corto puede existir un recuerdo de las respuestas dadas en la primera ocasión. En ambos casos se obtendrá una medida distorsionada de la repetibilidad. Además, algunos participantes pueden no aceptar que se administre el cuestionario en dos ocasiones, especialmente si es extenso.

Ejemplo 22.1. En la evaluación de la fiabilidad del cuestionario de salud *Nottingham Impact Profile (NIP),* uno de los aspectos que se tuvo en cuenta fue la repetibilidad (Hunt et al., 1981). El primer problema que se planteaba era la elección de la población. En estudios previos se había observado que las respuestas de los pacientes con enfermedades crónicas eran más fiables que las de los que padecían enfermedades agudas, y que las respuestas negativas eran más consistentes que las afirmativas. Dado que el NIP resulta en una alta proporción de respuestas negativas cuando se administra a una persona sana, una investigación sobre su fiabilidad se debería llevar a cabo en una población en la que se esperara un elevado número de respuestas afirmativas, con el fin de evitar su sobrevaloración. Por otro lado, para evitar una infraestimación de la repetibilidad, la muestra de la población debería padecer una enfermedad estable y que previsiblemente no cambiara en un corto espacio de tiempo. Por estas razones, seleccionaron a pacientes con artrosis. Aunque el estado físico de una persona con artrosis puede fluctuar diariamente, es muy poco probable que se produzcan cambios significativos en un período de semanas.

Los autores consiguieron reclutar a 73 personas que cumplían los criterios de inclusión, y a cada una se le envió un cuestionario, una carta explicando el propósito del estudio y un sobre libre de franqueo para la respuesta. A todos aquellos que respondieron se les envió un segundo cuestionario 4 semanas más tarde. Este período se eligió para minimizar la posible sobrevaloración de la repetibilidad debida al efecto del recuerdo de las respuestas efectuadas en la primera ocasión. Los autores obtuvieron una tasa de respuesta del 88% en el primer cuestionario y del 90% en el segundo, francamente alta para este tipo de estudios, y la repetibilidad del cuestionario, a su juicio, fue buena. Sin embargo, se ha de recordar que, en el sentido más estricto, estos resultados solo son aplicables a pacientes con artrosis. Muy posiblemente los resultados sean extrapolables a otras poblaciones sanas o con otras enfermedades, aunque esto debe ser evaluado en cada ocasión.

Fiabilidad interobservador

La evaluación de la fiabilidad interobservador consiste en estimar el grado de concordancia entre dos o más evaluadores (observadores). La demostración de una alta fiabilidad interobservador implica que la fiabilidad intraobservador también es alta. No obstante, si la fiabilidad interobservador es baja, no se puede asegurar si se debe a la existencia de diferencias entre los observadores o se debe a un solo observador.

Ejemplo 22.2. El *Older Americans Resources and Services Multidimensional*

Functional Assessment Questionnaire (OARS-MFAQ) es un cuestionario sobre la capacidad funcional y las necesidades de atención de las personas de edad avanzada. Dado que las puntuaciones del cuestionario se basan en una revisión de las respuestas por un observador, la fiabilidad interobservador es especialmente importante. En un estudio (Fillenbaum y Smyer, 1981) se evaluó la concordancia entre 11 observadores que evaluaron a 30 pacientes, y se obtuvieron coeficientes de correlación intraclase que oscilaron entre 0,66 para el estado físico y 0,87 para los autocuidados.

Consistencia interna

La consistencia interna se refiere a si los ítems que miden un mismo atributo presentan homogeneidad entre ellos.

Los cuestionarios se desarrollan para medir separadamente diferentes componentes o dimensiones de un problema. Un cuestionario de salud suele estar dividido en preguntas que tratan de medir la salud física y mental, o un cuestionario de satisfacción en apartados que identifican, por ejemplo, los componentes de competencia profesional, las cualidades personales del profesional sanitario, la información recibida, el trato y la accesibilidad de los servicios. En todas estas situaciones es de esperar que exista una buena homogeneidad entre las distintas preguntas que miden un mismo componente. Si en un cuestionario de satisfacción los usuarios contestan que su médico se preocupa bastante de ellos como persona, es de esperar que, en otra pregunta sobre si su médico está dispuesto a escucharlos, contesten afirmativamente. De otro modo, se pensará que los distintos ítems que componen la satisfacción con el médico que los atiende son poco consistentes entre sí y que el cuestionario es poco fiable.

Cuando un cuestionario está compuesto por diferentes subescalas, cada una de las cuales pretende medir una dimensión diferente del fenómeno, debe evaluarse la consistencia interna de cada una de ellas.

A diferencia de los otros aspectos de la fiabilidad, la evaluación de la consistencia interna solo requiere la administración del cuestionario en una única ocasión.

La técnica estadística para su análisis es el alfa de Cronbach, que expresa la consistencia interna entre tres o más variables. Sus valores están comprendidos entre 0 y 1, y su interpretación es similar a la de un coeficiente de correlación. Pueden calcularse diferentes valores del alfa de Cronbach excluyendo determinados ítems del cuestionario, de forma que puede evaluarse si la supresión de algunas preguntas mejora la fiabilidad. De todas formas, antes de decidir eliminar un ítem, debe evaluarse si ello puede afectar a la validez del cuestionario, ya que, debido a su importancia, puede ser preferible mantener la pregunta aun a costa de una consistencia interna ligeramente menor.

Como norma general, se sugiere que el valor del alfa de Cronbach ha de ser igual o superior a 0,70 para considerar que un instrumento tiene una buena consistencia interna.

Ejemplo 22.3. El cuestionario de apoyo social funcional Duke-UNC-11 consta de 11 ítems medidos en una escala de Likert con puntuaciones de 1 a 5, y evalúa el apoyo social confidencial (posibilidad de contar con personas para comunicarse con ellas) y el afectivo (demostraciones de amor, cariño y empatía). En el marco de un estudio de su validez y fiabilidad, se evaluó también su consistencia interna (Bellón et al, 1996). El alfa de Cronbach de la escala fue de 0,90, el de la subescala de apoyo confidencial de 0,88 y el del apoyo afectivo de 0,79. Los autores no observaron que la extracción de ningún ítem mejorara la consistencia interna de la escala ni de las subescalas.

Fuentes de error

La fiabilidad de una medida puede verse afectada por algunos de los factores que se citan a continuación:

- *Cambios a través del tiempo en la característica estudiada.* Al repetir un cuestionario se debe tener en cuenta qué medidas pueden variar con el tiempo. De hecho, muchas actitudes, creencias o estilos de vida pueden hacerlo, como la frecuencia de cepillado de los dientes o la práctica de ejercicio físico. Si ha existido un

cambio, una repetibilidad baja no implica necesariamente una escasa fiabilidad del cuestionario.

- *Cambios debidos a las condiciones de administración del cuestionario.* Algunos factores personales del entrevistado, como el estado emocional, el cansancio, el estado de salud o las condiciones del entorno (ruido, calor, frío, etc.), pueden influir en el modo de contestar a las preguntas y alterar la fiabilidad.
- *Variaciones debidas al propio cuestionario.* En ocasiones, el formato del cuestionario, la formulación de las preguntas o las instrucciones para cumplimentarlo son poco comprensibles y pueden ser interpretadas de forma distinta por el encuestado al repetir la prueba. Por consiguiente, hay que insistir en que las instrucciones sean cortas y precisas y que las preguntas estén definidas de forma operativa con un lenguaje claro y sin ambigüedades.
- *Cambios atribuidos a los encuestadores.* Los encuestadores deben atenerse estrictamente a la estructura y la secuencia del cuestionario, y a cómo han sido formuladas las preguntas. La transcripción de las respuestas a las preguntas abiertas debe ser literal. Al poner en marcha un estudio es de suma importancia entrenar a los entrevistadores para evitar estos errores y conseguir al mismo tiempo que exista uniformidad entre ellos, lo que suele implicar formular exactamente las preguntas tal como han sido elaboradas, sin explicaciones adicionales (no previstas) y leyendo todas las opciones de respuesta.
- *Errores en el manejo de los datos.* Estos errores se pueden producir al codificar, grabar o transformar las variables para su análisis.

VALIDEZ

La validez se refiere a la capacidad de un cuestionario para medir aquello para lo que ha sido diseñado; tiene diferentes aspectos o componentes, que deben ser evaluados en la medida de lo posible.

Validez lógica

La validez lógica o aparente *(face validity)* es el grado en que parece que un cuestionario, una parte de él o un ítem mide lo que quiere medir.

La decisión sobre si las preguntas deben tener o no validez lógica ha de tomarse antes de iniciar su redacción. Si las preguntas carecen de validez lógica es muy probable que los encuestados rechacen contestar. De todos modos, en alguna ocasión puede ser de interés formular preguntas carentes de validez lógica. Por ejemplo, cuando se trata de temas muy sensibles, conflictivos o que no están bien vistos socialmente, si se realizan preguntas directas (con mucha validez lógica) es muy probable que el encuestado no responda o falsee la respuesta, por lo que puede ser preferible realizar preguntas que aborden el tema de una forma más indirecta, con una menor validez aparente.

Validez de contenido

La validez de contenido se basa en el análisis del concepto que se pretende medir y, en especial, en la definición de las áreas o dimensiones que abarca y sus límites con otros conceptos relacionados. Se puede considerar que un cuestionario es válido por su contenido si contempla todos los aspectos relacionados con el concepto en estudio, lo que suele evaluarse a través de la opinión de expertos.

Ejemplo 22.4. Supongamos que se desarrolla un cuestionario para evaluar los conocimientos que tienen los estudiantes de medicina sobre una determinada enfermedad o sobre varios grupos de enfermedades. Para determinar la validez de contenido podría construirse una tabla en la que las columnas representarían las diferentes áreas de conocimiento (anatomía, fisiología, etiología, diagnóstico, etc.) y en cada columna se señalarían las preguntas del cuestionario que se incluyeran en dicha área. La simple inspección visual de la tabla permitiría apreciar si todas las áreas de conocimiento están representadas en la proporción deseada. El número de preguntas de cada área dependería de la importancia relativa de su

contenido, habitualmente determinada por las opiniones de expertos e investigaciones previas sobre el tema.

Una forma empírica de evaluar la validez de contenido es aplicar un análisis factorial, técnica estadística que explora las respuestas a las preguntas del cuestionario, intentando agruparlas en función de factores subyacentes. Por ejemplo, si un cuestionario de estado de salud contiene dos subescalas, una que mide la salud física y otra la mental, es de esperar que el análisis detecte dos factores, cada uno de ellos formado por las preguntas que se relacionan con cada una de las subescalas. Para aplicar el análisis factorial, las escalas de medida deben ser cuantitativas o de puntuación por intervalos, y las respuestas han de seguir una distribución aproximadamente normal.

Ejemplo 22.5. En el estudio presentado en el ejemplo 22.3, en el que se evaluaba el cuestionario de apoyo social funcional Duke-UNC-11, que consta de 11 ítems medidos en una escala de Likert con puntuaciones de 1 a 5, se realizó un análisis factorial para detectar las dimensiones subyacentes en el cuestionario. El análisis reveló la existencia de dos factores: el primero con 7 ítems y el segundo con 4 ítems. Ambos factores explicaban, conjuntamente, el 60,9% de la variabilidad. Estos factores corresponden a los dos componentes teóricos del cuestionario original: el apoyo social confidencial (posibilidad de contar con personas para comunicarse con ellas) y el afectivo (demostraciones de amor, cariño y empatía).

La diferencia entre la validez aparente y la de contenido reside en que la evaluación de esta última es un proceso más exhaustivo, y quizá más formal, y en el que deberían participar tanto investigadores y clínicos como miembros de la población diana.

Validez de criterio

En ocasiones se puede disponer de algún método alternativo de medida del fenómeno estudiado cuya validez haya sido demostrada, que se toma como referencia para determinar la validez de la encuesta. Siempre que se disponga de un método de referencia adecuado, deberá evaluarse la validez de criterio del cuestionario.

Cuando se habla de validar un cuestionario, los investigadores se suelen referir a la validez de criterio, que es, sin lugar a dudas, la más importante. En algunos casos se pueden usar como criterio de referencia medidas bioquímicas o radiológicas. Se puede validar, por ejemplo, el consumo de tabaco declarado, comparándolo con los valores derivados de la nicotina en sangre o de monóxido de carbono en el aire espirado. En otros casos, el investigador tendrá que fiarse de medidas menos objetivas, como la historia clínica o los resultados obtenidos mediante otro cuestionario.

La validez de criterio puede evaluarse de dos formas: la validez concurrente y la validez predictiva. Para valorar la *validez concurrente* se relaciona la nueva medida con la de referencia, ambas administradas simultáneamente, de forma similar al estudio de la utilidad de una prueba diagnóstica. Cuando el criterio de referencia no esté disponible hasta un tiempo después (p. ej., el desarrollo de una enfermedad), se valora hasta qué punto la nueva medida es capaz de predecirlo de manera correcta, y se habla de *validez predictiva*. Cuando se evalúa la validez predictiva, los resultados del cuestionario no pueden influir sobre el criterio de referencia. En otras palabras, la evaluación de este criterio de referencia debe hacerse independientemente del resultado del cuestionario.

Ejemplo 22.6. Dado que la relación entre la capacidad pulmonar y la calidad de vida en enfermos pulmonares crónicos es débil, Guyatt et al. (1987) desarrollaron un cuestionario sobre calidad de vida para utilizarlo en ensayos clínicos. Los existentes hasta el momento se centraban en la disnea y dejaban a un lado otros aspectos que pueden influir en la vida cotidiana de los pacientes. Por ello, desarrollaron un cuestionario que comprende cuatro grandes dimensiones: disnea, fatiga, estado emocional y la sensación de control del propio paciente sobre la enfermedad. La repetibilidad, que se evaluó en 100 pacientes con limitación crónica al flujo aéreo estable, fue excelente para las cuatro dimensiones. Los autores también evaluaron la validez de criterio,

comparando las puntuaciones del cuestionario con los resultados espirométricos y con otros cuestionarios destinados a medir la disnea y la calidad de vida en general.

Cuando la escala de medida es cualitativa, los índices que se utilizan para evaluar la validez de criterio son la sensibilidad y la especificidad (anexo 3). Cuando se trata de una escala cuantitativa, se utiliza habitualmente el coeficiente de correlación intraclase (anexo 4).

A menudo surge el interrogante de por qué, si ya existe un buen criterio de referencia, interesa una nueva medida. El desarrollo de esta nueva medida está justificado si el criterio de referencia es muy caro, requiere mucho tiempo de administración, presenta muchos efectos secundarios, o bien no se desarrolla hasta un tiempo después. En las tres primeras situaciones el interés se centra en evaluar la validez concurrente, mientras que en la última interesa determinar la validez predictiva del cuestionario.

Validez de constructo o de concepto

A veces resulta imposible evaluar la validez de criterio, ya que este no existe o no está al alcance del investigador. En estos casos, el procedimiento más empleado es evaluar la validez de constructo, que engloba distintas estrategias. La *validez discriminante* se refiere a la capacidad para distinguir entre subgrupos de pacientes o individuos con distintos niveles del atributo de interés. Por ejemplo, es de esperar que la calidad de vida relacionada con la salud sea peor en los pacientes asmáticos con gran afectación funcional que en aquellos con formas más leves.

El método más sencillo para evaluar la validez discriminante es el de los *grupos extremos,* que consiste en administrar el cuestionario a dos grupos de sujetos: uno con la característica o conducta de interés, y otro que carece de ella. Este enfoque presenta dos problemas. El primero es la propia definición de los grupos extremos, ya que no siempre existe un criterio adecuado para conocer quién tiene y quién no tiene la característica de interés. En este caso se puede dividir la muestra en función de la puntuación obtenida con el propio instrumento, seleccionando, por ejemplo, el 30% de los sujetos con mejores puntuaciones y el 30% de los que tienen las peores puntuaciones. El segundo problema es similar al que se presenta al evaluar una prueba diagnóstica: puede ser relativamente sencillo discriminar entre dos grupos muy extremos, pero esta no es la utilidad que se pretende dar al instrumento en la práctica habitual. Por tanto, comprobar que un cuestionario es útil para diferenciar entre dos grupos extremos no es suficiente para demostrar su validez.

Ejemplo 22.7. En un estudio que tenía por objetivo evaluar la validez y fiabilidad de un cuestionario de función familiar (Bellón et al., 1996) se utilizó el método de los grupos extremos para evaluar la validez de constructo. Se partió de la hipótesis de que los casados tienen una mejor función familiar que los divorciados, por lo que deberían obtener puntuaciones superiores en el cuestionario.

Otra estrategia para evaluar la validez de constructo es comprobar que el cuestionario se correlaciona con otras variables que se cree que están relacionadas con él *(validez convergente),* mientras que no lo hace con otras con las que se sospecha que no tiene relación alguna *(validez divergente).*

Ejemplo 22.8. Volviendo al estudio que evaluó la validez y la fiabilidad del cuestionario de apoyo social funcional Duke-UNC-11 en una muestra de 656 pacientes (Bellón et al., 1996), para evaluar la validez de constructo los autores eligieron determinadas características que, según la literatura científica, están relacionadas con el apoyo social, y que son la edad, el estado civil, el estado de salud, la utilización de servicios, la salud mental, la función familiar y el número de convivientes. Posteriormente compararon las distribuciones de estas variables entre los sujetos con un apoyo social normal o bajo, según el cuestionario, y determinaron los coeficientes de correlación entre el cuestionario evaluado y el resto de las escalas cuantitativas.

Ejemplo 22.9. Para validar el test de Fresno, un cuestionario que mide los

conocimientos y habilidades de los profesionales sanitarios en medicina basada en la evidencia (Argimon-Pallàs et al., 2010), los autores seleccionaron tres grupos en función de su experiencia previa. El primero estaba formado por profesionales expertos que habitualmente participaban en el diseño y la realización de ensayos clínicos, el segundo por tutores de medicina de familia, y el tercero por médicos residentes. La hipótesis subyacente era que la puntuación del test sería mayor en los grupos más expertos comparada con la de los residentes. El grupo con profesionales más expertos obtuvo una puntuación de 149,8 sobre un máximo de 212 puntos, mientras que los residentes solo obtuvieron 60,4 puntos. Los tutores de medicina de familia obtuvieron una puntuación intermedia (110,4 puntos), lo que confirmó la hipótesis de los investigadores.

SENSIBILIDAD AL CAMBIO

La sensibilidad al cambio *(responsiveness)* es la capacidad que tiene un instrumento para detectar cambios importantes en el atributo que se mide a lo largo del tiempo, y puede considerarse un tipo especial de validez de constructo o concepto. La sensibilidad al cambio es un aspecto crucial en los ensayos clínicos, la valoración de programas y los análisis de coste-utilidad; en otras palabras, cuando el instrumento se utiliza como variable de respuesta.

Su conocimiento también es importante para la estimación del cálculo del tamaño muestral necesario para demostrar los cambios debidos al tratamiento en un ensayo clínico. Cuanto mayor sea la sensibilidad al cambio del instrumento, menor será el tamaño de la muestra necesario.

La sensibilidad al cambio se estima en estudios longitudinales y existen dos abordajes para calcularla. El primero se basa en la distribución estadística y evalúa los cambios que se producen en la puntuación de la escala y su variabilidad asociada, generalmente la desviación estándar. Existen varios índices que la miden, todos ellos muy parecidos. Los índices más usados son el tamaño del efecto *(effect size)*, la media del cambio estandarizada *(standardised response mean)* y el índice de sensibilidad al cambio.

Aunque hay diferentes fórmulas para calcularlo, en líneas generales el tamaño del efecto se corresponde con la diferencia entre los valores basales y los del final del estudio (magnitud del cambio) dividida por una medida de dispersión, normalmente la desviación estándar de los valores basales. Su resultado refleja el cambio estandarizado en el número de desviaciones estándar de los valores basales. Como norma general, el tamaño del efecto puede considerarse pequeño si es menor de 0,2, pequeño-moderado si se encuentra entre 0,2 y 0,5, moderado-grande si va de 0,51 a 0,79, y grande si supera el valor de 0,79.

La media del cambio estandarizada se calcula como la media del cambio en las puntuaciones, antes y después, dividida por la desviación estándar de estos cambios. Los valores positivos reflejan incrementos estandarizados en el número de desviaciones estándar de las diferencias de puntuaciones.

El índice de sensibilidad al cambio se define como la diferencia media entre los valores basales y los obtenidos al final del seguimiento, dividida por la desviación estándar de los individuos más estables (p. ej., aquellos que han recibido un placebo). Este índice es el que suele mostrar valores más elevados, ya que la desviación estándar es más pequeña.

La ventaja de estos índices estadísticos es que son fáciles de calcular, aunque presentan dos limitaciones importantes: los valores de dispersión (variabilidad) difieren de un estudio a otro y su interpretación no es muy intuitiva para el clínico. Además, los cambios «estadísticamente significativos» dependen del número de participantes en el estudio y, por tanto, no equivalen necesariamente a un cambio clínicamente importante, o viceversa.

Ejemplo 22.10. En un estudio se incluyeron 152 médicos residentes y se analizaron los cambios en sus conocimientos sobre medicina basada en la evidencia, medidos con el test de Fresno, antes y después de recibir un seminario (Argimon-Pallás et al., 2011). El tamaño del efecto fue de 1,77 (intervalo de confianza del 95% [IC 95%]: 1,57-1,95)

y la respuesta media estandarizada fue de 1,65 (IC 95%: 1,47-1,82). Las diferencias entre ambos índices, aunque no son muy grandes, se explican porque persiguen objetivos distintos. Mientras que los dos índices usan la diferencia entre las puntuaciones obtenidas antes y después de la intervención, el estimador de la variabilidad es distinto. La respuesta media estandarizada usa la desviación estándar de las diferencias y su objetivo es mostrar que existen cambios estadísticamente significativos, mientras que el tamaño del efecto usa la desviación estándar antes de la intervención y su objetivo es cuantificar la magnitud del efecto.

En el segundo abordaje para la determinación de la sensibilidad al cambio, denominado método de ancla, las puntaciones de la escala que se valida se comparan con los valores de otra variable o escala validada, con el fin de observar si las puntuaciones de los dos instrumentos se correlacionan. La sensibilidad al cambio estimada por este método siempre es más sencilla de interpretar. Lo que es importante es que la variable que se usa como ancla sea por sí misma una medida válida del cambio clínico que puedan experimentar los participantes del estudio. Su principal limitación es que no tiene en cuenta ninguna medida de la precisión de la variable, y de este modo no se puede descartar que un cambio relevante en las puntuaciones sea debido a una gran variabilidad de la medida.

Para que una escala pueda emplearse como variable de respuesta en un ensayo clínico, es necesario que los ítems que la forman midan todos los conceptos relevantes de la enfermedad o del problema que se esté estudiando, y que la mayoría de ellos sean capaces de detectar cambios. La sensibilidad al cambio vendrá determinada por la interacción entre los propios ítems que formen la escala, la intervención que se evalúe y la población en la que se aplique.

Si al contestar un ítem de un cuestionario los pacientes no reflejan tener un problema de salud, entonces este ítem no será válido para detectar una mejoría. Del mismo modo, si un ítem solo se altera en los casos más graves, en los estudios donde se incluyan pacientes menos graves este ítem tampoco será útil para detectar cambios. Por otro lado, si un ítem se altera en todos los pacientes porque es una característica de la enfermedad, tampoco servirá para detectar cambios, a menos que la intervención que se evalúe consiga la curación del paciente. En definitiva, los ítems que muestren un efecto «techo» *(ceiling effect)* o un efecto «suelo» *(floor effect)* no serán sensibles para detectar cambios.

Los efectos «techo» o «suelo» dependen, en parte, de la población de estudio. Un ítem puede mostrar un efecto «cielo» cuando se estudia a pacientes muy graves, pero no cuando estos pacientes tienen un grado menor de afectación del estado de salud. Los efectos «cielo» o «techo» se pueden inferir de los resultados de las pruebas piloto o de los estudios que se hayan efectuado en distintas poblaciones.

Estos efectos también dependen de la validez de contenido del cuestionario. Si un ítem es irrelevante para un grupo de población, su capacidad para detectar un cambio a lo largo del estudio será pequeña.

Ejemplo 22.11. Los resultados de algunas investigaciones sobre el asma sugieren que los ítems relacionados con la práctica de un deporte son más sensibles al cambio en las personas jóvenes, mientras que en las mayores lo son menos. Una explicación es que, para las personas mayores, la posibilidad de practicar un deporte es menos importante. La relevancia de un ítem depende mucho de la población en la que se administra el cuestionario.

La relevancia de un ítem se manifiesta al comparar escalas específicas de una enfermedad o un problema de salud para medir la calidad de vida o el estado de salud percibido por los pacientes con escalas genéricas. Como norma general, las escalas específicas suelen contener más ítems con capacidad para detectar cambios que las genéricas y, por consiguiente, son las más usadas en los ensayos clínicos.

El que un ítem sea o no sensible a los cambios también depende del formato de respuesta. En general, las respuestas tipo escala de Likert son las más sensibles para detectar cambios. Cuantos más puntos haya en las

categorías, más sensible será el instrumento a los cambios.

Fuentes de error

Los aspectos del diseño del cuestionario que pueden influir en su validez son los siguientes:

- *Orden de las preguntas.* Es conveniente situar las más conflictivas al final del cuestionario, ya que si se formulan al principio, existe la posibilidad de que la persona encuestada rechace seguir respondiendo o no lo haga con la sinceridad deseada.
- *Redacción de las preguntas,* que puede inducir una respuesta sesgada.

Ejemplo 22.12. Supongamos una pregunta en la que se plantea la posibilidad de que una enfermedad afecte a 600 individuos y se informa de que existen dos programas alternativos: el programa A, que salvará 200 vidas, y el B, con el que hay un tercio de probabilidades de salvarse y dos tercios de morir. La gran mayoría de los encuestados preferirá el programa A. Nótese que los resultados de los programas A y B (en términos de vidas salvadas) son los mismos. La diferencia está en cómo se ha explicado o descrito la situación. En el programa A, el número de supervivientes queda explícito, mientras que el número que muere (400) queda implícito. Desde un punto de vista aritmético, el número de sujetos que debería preferir uno u otro programa tendría que ser el mismo o similar. Esta aparente paradoja se puede explicar porque los individuos tienen aversión al riesgo cuando se presenta una situación de posible ganancia y, por el contrario, aceptan el riesgo cuando la situación se presenta desde el punto de vista de pérdida.

- *Errores en la categorización de las respuestas.* En los cuestionarios con opciones de respuestas predeterminadas hay que tener en cuenta todas las posibilidades de respuesta, de modo que el encuestado encuentre siempre una opción adecuada. Por ello, es importante que el entrevistador formule todas las opciones antes de registrar la respuesta. En otras ocasiones, son los factores personales del entrevistado los que introducen los sesgos en las respuestas. Hay individuos que, al ser encuestados, tienden a contestar de forma distinta a como lo harían normalmente.
- *Sesgo de conveniencia social.* Algunas personas tienden a dar la respuesta socialmente más aceptable o la que piensan que contesta la mayoría de los individuos. Este sesgo depende de muchos factores individuales, como la edad, el sexo, la clase social del entrevistado y el contexto en el que se realiza la pregunta. Ejemplos de ello se pueden encontrar en cuestionarios o preguntas sobre el consumo de alcohol o el aborto. Un mecanismo para intentar evitarlo es dar en primer lugar las opciones menos convenientes, de modo que sea más fácil seleccionarlas. En muchas ocasiones, el sesgo de conveniencia social se produce sin que el entrevistado tenga intención de engañar, mientras que en otras el individuo no dice lo que piensa de forma deliberada.
- *Tendencia sistemática a dar siempre la misma respuesta.* Otra fuente de error es la tendencia sistemática de algunos individuos a contestar «sí», «verdadero» o «de acuerdo» a las cuestiones planteadas. En el caso más extremo, estas respuestas se dan independientemente del contenido de la pregunta. Como ejemplo se podría dar el caso de un individuo que contestara afirmativamente a la pregunta de si toma medicación todos los días a la hora indicada, y también lo hiciera cuando se le preguntara si a menudo se olvida de tomar la medicación. En el otro extremo del espectro se encuentran los que siempre contestan negativamente o no están de acuerdo.
- *Características del encuestador.* La respuesta de una persona a la pregunta de si cumple con la medicación prescrita puede ser muy distinta si la realizan los profesionales sanitarios que atienden al paciente o si lo hace un entrevistador no implicado directamente en su seguimiento médico.
- *Elección de la categoría* o *puntuación intermedia.* En las escalas de puntua-

ción o las de Likert existe la posibilidad, además, de que aparezcan otros sesgos. Uno de ellos se refiere a que algunos individuos siempre escogen la categoría o puntuación intermedia. El efecto de este sesgo es la reducción, en la práctica, de las posibles respuestas. Así, si en una escala de Likert con cinco opciones, las extremas recogen la posibilidad de «siempre» o «nunca», un *sesgo de aversión hacia los extremos* implica que la escala queda reducida a tres categorías, ya que el individuo no contesta «nunca» ni «siempre», con la consiguiente pérdida de fiabilidad. Existen dos posibles soluciones a este problema. La primera es evitar términos absolutos y, en su lugar, utilizar «casi siempre» y «casi nunca». La segunda consiste en aumentar las categorías posibles en cada respuesta; así, si en realidad se desea una pregunta con cinco categorías de respuesta, finalmente tendrá siete, de modo que las categorías extremas sean las ocupadas por los términos absolutos.

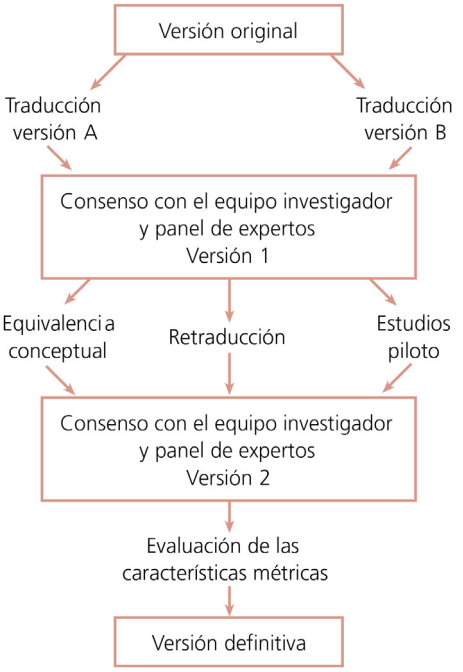

FIGURA 22.1 Esquema del proceso de adaptación transcultural de un cuestionario.

ADAPTACIÓN TRANSCULTURAL

La adaptación de un cuestionario a otra cultura o entorno tiene como objetivo conseguir un instrumento equivalente al desarrollado en el país de origen. No puede limitarse a una simple traducción, sino que debe seguir una metodología que asegure la equivalencia conceptual y semántica con el original y la comprensión por parte de los pacientes de la versión adaptada. El método más utilizado es el de la *traducción-retraducción* por personas bilingües (fig. 22.1), seguido de un análisis de la nueva versión para detectar discrepancias, y de la comprobación de su comprensión y aceptabilidad en un grupo de pacientes.

Para realizar una adaptación transcultural, en primer lugar debe evaluarse si el fenómeno que tiene que medir el cuestionario existe en la cultura a la que se desea adaptar. Por ejemplo, si el cuestionario trata sobre cómo una enfermedad o un problema de salud afecta al grado o la manera de realizar una determinada actividad, hay que preguntarse si estos conceptos se operativizan de igual modo en una cultura que en otra. Es esencial conocer qué significado tiene el concepto y cómo se expresa en la cultura original para poder identificar similitudes en la nuestra.

A continuación, hay que traducir el cuestionario. Deben realizarse al menos dos traducciones conceptuales más que literales. La persona que traduce el cuestionario no solo debe ser un perfecto conocedor de los dos idiomas, sino también de los contenidos y los propósitos del cuestionario, ya que la traducción literal de un término puede tener un significado distinto según el idioma. Por ejemplo, en inglés se usa la palabra *blue* para designar tristeza. Si la frase *«I feel blue»* se tradujera literalmente al español no tendría ningún sentido. A partir de estas traducciones, el equipo investigador consensua una primera versión del cuestionario. Es conveniente que un panel de expertos evalúe la equivalencia conceptual de esta versión con la original.

Después se retraduce el cuestionario dos veces al idioma original por al menos dos in-

dividuos bilingües, evaluando su equivalencia conceptual con el cuestionario original. Es conveniente realizar una revisión del cuestionario por un panel de población general o de pacientes de diferentes características sociodemográficas con la finalidad de llegar a un acuerdo sobre los ítems que se van a incluir. A partir de toda esta información se consensúa una versión preliminar. Finalmente, debe realizarse un estudio piloto de esta versión en una muestra de sujetos de características similares a las de la población a la que se administrará el cuestionario, para poder evaluar su comprensión y viabilidad.

Una vez obtenida la versión final, deben comprobarse sus propiedades métricas mediante el reescalamiento de los ítems y dimensiones en el nuevo medio, ya que su importancia, o valor relativo, puede ser distinta en ambas culturas. Y, finalmente, deben comprobarse su validez y fiabilidad. Además, si pretende utilizarse para medir la respuesta en estudios longitudinales, hay que evaluar su sensibilidad al cambio.

Ejemplo 22.13. Un estudio tenía por objetivo adaptar el cuestionario genérico de estado de salud *SF-36 Health Survey* para su uso en España (Alonso et al., 1995). Este proyecto se estaba realizando de forma coordinada en diferentes países. El cuestionario original fue traducido al castellano por dos personas bilingües con experiencia clínica. Ambas traducciones fueron discutidas simultáneamente entre los traductores y un miembro del equipo investigador hasta alcanzar un consenso. Otras dos personas bilingües evaluaron la equivalencia conceptual, la claridad y la naturalidad de cada una de las frases y de las opciones de respuesta de esta primera versión. A continuación fue traducido al inglés por dos personas bilingües. Sus dos retraducciones fueron comparadas con la versión original por un equipo de expertos, quienes señalaron algunos ítems o palabras que no parecían tener una completa equivalencia conceptual con el original. También se realizó una reunión con los autores de todas las versiones del cuestionario existentes en diferentes países, durante la cual se trató de armonizar el contenido del cuestionario en los casos en que existían diferentes expresiones alternativas del mismo concepto. Por ejemplo, se sustituyó la distancia de una milla por la de un kilómetro. Se realizaron diferentes estudios piloto con diversos grupos de pacientes crónicos para valorar la comprensibilidad del cuestionario y la factibilidad de su administración.

SELECCIÓN Y USO DE UN INSTRUMENTO DE MEDIDA

Una vez validados, los instrumentos de medida se usan generalmente con dos fines distintos: para evaluar la efectividad de una intervención sanitaria, o para describir o discriminar entre grupos de pacientes.

Las escalas que se usan en los ensayos clínicos para valorar la efectividad han de contener una elevada proporción de ítems sensibles al cambio, no suelen ser muy largas y, habitualmente, no superan los 30 o 40 ítems.

Por el contrario, cuando un instrumento se emplea en un estudio transversal con el fin de describir el estado de salud o discriminar entre los pacientes, no es necesario que el número de ítems sea reducido. Un instrumento tendrá más capacidad de discriminación cuanto mayor sea el número de ítems que contenga.

En un ensayo clínico no conviene usar instrumentos que contengan ítems frente a los que más del 70% de los pacientes opte por la respuesta más extrema (efecto «techo» o «cielo»). Sin embargo, en un estudio transversal interesa usar instrumentos con los que los pacientes opten por todas las posibles respuestas en función de su gravedad. Por tanto, los pacientes más graves o con menos calidad de vida optarán por una respuesta extrema, mientras que los que tengan menor afectación optarán por otra respuesta.

La elección entre una escala genérica o específica para una enfermedad dependerá del objetivo del estudio.

Ejemplo 22.14. Supóngase que en un estudio se emplea un instrumento genérico para valorar la calidad de vida de un grupo de pacientes diagnosticados de enfermedad pulmonar obstructiva crónica (EPOC), y que este instrumento contiene un número importante de ítems relacionados con el dolor.

Como los pacientes con EPOC suelen ser mayores y presentan otras enfermedades, algunas de ellas cuyo síntoma principal puede ser el dolor, existirá una gran variación en las puntuaciones de estos enfermos. Supóngase que se desea, a continuación, correlacionar la puntuación obtenida en el cuestionario de calidad de vida con una medida de la función pulmonar. Debido a la gran variabilidad observada en la puntuación del cuestionario de calidad de vida, es posible que se reduzca la posible asociación entre calidad de vida y la medida de la función pulmonar. En este caso sería preferible el uso de un cuestionario específico, en el que no existieran interferencias de preguntas que no tuvieran una relación directa con la enfermedad. Por el contrario, si se deseara una medida general del estado de salud de una población, o compararlo con el de otro grupo, teniendo en cuenta la comorbilidad, sería preferible una medida genérica.

La utilidad de un instrumento depende básicamente de su fiabilidad y validez, pero también de su interpretabilidad clínica. Esta viene dada por el grado en que se pueden realizar juicios de valor sobre un resultado cuantitativo, que permitan, por ejemplo, la toma de decisiones clínicas. La estrategia más utilizada para aumentar la interpretabilidad de los instrumentos de medida de la salud percibida ha sido su administración a una muestra representativa de la población general para obtener valores o normas poblacionales de referencia, basados habitualmente en el cálculo de los percentiles.

Cuando el instrumento se usa como variable de respuesta, además de la sensibilidad al cambio es importante disponer del valor que puede considerarse como el cambio mínimo que puede estimarse como relevante desde el punto de vista clínico. El método más recomendado es usar varios marcadores clínicos o anclas, junto con el tamaño del efecto o la respuesta media estandarizada como información de soporte, para finalizar obteniendo el valor o rango de valores considerados relevantes desde el punto de vista clínico.

En ausencia de una medida objetiva o ancla, la diferencia mínima de interés clínico puede estimarse como una proporción la desviación estándar del tamaño del efecto.

Ejemplo 22.15. Norman et al. (2003) realizaron una revisión sistemática de 38 estudios (que incluían 62 tamaños del efecto) y observaron, con solo unas pocas excepciones, que las mínimas diferencias clínicamente relevantes se correspondían con la mitad de la desviación estándar del tamaño del efecto.

Es preferible usar la desviación estándar de la puntuación basal del tamaño del efecto antes que la desviación estándar de los cambios antes y después, como en la respuesta media estandarizada, ya que el propósito es decidir la magnitud del cambio, no su significación estadística.

De todos modos, la diferencia mínima no debe considerarse un valor fijo, sino que depende del contexto y puede variar de un estándar a otro y entre poblaciones.

BIBLIOGRAFÍA DE LOS EJEMPLOS

Alonso J, Prieto L, Antó JM. La versión española del SF-36 Health Survey (Cuestionario de Salud SF-36): un instrumento para la medida de los resultados clínicos. Med Clin (Barc). 1995;104:771-6.

Argimon-Pallás JM, Flores-Mateo G, Jiménez-Villa J, Pujol-Ribera E. Effectiveness of a short-course in improving knowledge and skills on evidence-based practice. BMC Fam Pract. 2011;12:64.

Argimon-Pallás JM, Flores-Mateo G, Jiménez-Villa J. Pujol-Ribera E. Psychometric properties of a test in evidence based practice: the Spanish version of the Fresno test. BMC Med Educ. 2010;10:45.

Bellón JA, Delgado A, Luna del Castillo JD, Lardelli P. Validez y fiabilidad del cuestionario de apoyo social funcional Duke-UNC-11. Aten Primaria. 1996;18:153-63.

Bellón JA, Delgado A, Luna del Castillo JD, Lardelli P. Validez y fiabilidad del cuestionario de función familiar Apgar-familiar. Aten Primaria. 1996;18:289-96.

Fillenbaum GG, Smyer MA. The development, validity and reliability of the OARS Multidimensional Functional Assessment Questionnaire. J Gerontol. 1981;36:428-34.

Guyatt G, Berman L, Towsend M, Pugsley S, Chambers L. A measure of quality of life for clinical trials in chronic lung disease. Thorax. 1987;42:773-8.

Hunt S, Mc Kenna, Williams J. Reliability of a population survey tool for measuring perceived health problems: a study of patients with osteoarthrosis. J Epidemiol Community Health. 1981;35:297-300.

Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care. 2003;41:582-92.

BIBLIOGRAFÍA

Altman DG, Bland JM. Cronbach's alpha. BMJ. 1997;314:572.

Badia X, Salamero M, Alonso J. La medida de la salud: guía de escalas de medición en español. 4.ª ed. Barcelona: Unión Editorial; 2007.

Bland JM, Altman DG. Validating scales and indexes. BMJ. 2002;324:606-7.

McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. 3rd ed. New York: Oxford University Press; 2006.

Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008;61:102-9.

Patten M. Questionnaire design: a practical guide. 4th ed. New York: Routledge; 2017.

Streiner D, Norman G, Cairney J. Health measurement scales – a practical guide to their development and use. 5th ed. New York: Oxford University Press; 2015.

Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res. 2003;12:349-62.