Análisis factorial confirmatorio. Su utilidad en la validación de cuestionarios relacionados con la salud

Joan Manuel Batista-Foguet[a], Germà Coenders[b] y Jordi Alonso[c]

[a]*ESADE. Dirección de Investigación. Universitat Ramon Llull. Barcelona.*
[b]*Institut de Recerca sobre Qualitat de Vida. Universitat de Girona. Girona. España.*
[c]*Unidad de Investigación en Servicios Sanitarios. Institut Municipal d'Investigació Mèdica (IMIM-IMAS). Barcelona. España.*

El investigador que usa cuestionarios en ciencias de la salud se basa tácitamente (a menudo, inadvertidamente) en la teoría clásica del test, cuyos supuestos, poco realistas, se incumplen con frecuencia y conducen a evaluar deficientemente la fiabilidad y validez del instrumento.
Este artículo destaca la necesidad de precisar los términos esenciales de la medición (fiabilidad y validez) clarificando las deficiencias en que incurre la metodología tradicional en su uso. Señala las limitaciones de la evaluación de la fiabilidad mediante el α de Chronbach o la laxitud en la valoración cuantitativa de la validez mediante el análisis factorial exploratorio.
Como alternativa se propone el tratamiento secuencial e integrado de la validez y la fiabilidad en el marco de los modelos del análisis factorial confirmatorio. Estos modelos proporcionan el marco estadístico adecuado para evaluar la validez y la fiabilidad de cada ítem, en lugar de efectuar sólo valoraciones globales. La perspectiva confirmatoria guía al investigador para que optimice el proceso de construcción o adaptación de un cuestionario, liberándole del ritual poco fundamentado que «recetaba» la metodología clásica.

Palabras clave: Validez. Fiabilidad. Análisis factorial confirmatorio. Análisis factorial exploratorio. α de Chronbach.

Confirmatory factor analysis. Its role on the validation of health related questionnaires

Researchers who use questionnaires in the health sciences tacitly base themselves (often inadvertently) in Classical Test Theory, the suppositions of which are unrealistic and frequently violated, leading to defective evaluation of the reliability and validity of the instrument.
The present article emphasizes the need for precise definition of essential terms of measurement (reliability and validity) by clarifying the deficiencies that traditional methodology incurs in their use. The limitations of evaluation of reliability through Chronbach's alpha and laxity in quantitative evaluation of validity through exploratory factor analysis are described.
As an alternative, a sequential and integrated approach to validity and reliability within the framework of confirmatory factor analysis models is proposed. These models provide the appropriate statistical framework to evaluate the validity and reliability of each item, instead of carrying out overall evaluations only. The confirmatory approach guides researchers to optimize the process of designing or adapting a questionnaire, freeing them from the largely unfounded ritual laid down by classical methodology.

Key words: Validity. Reliability. Confirmatory factor analysis. Exploratory factor analysis. Chronbach's α.

Introducción

En medicina, las variables de interés son frecuentemente de naturaleza no observable. La discapacidad, la calidad de vida o el resultado de una intervención son conceptos y abstracciones que representan fenómenos que sirven para comunicarnos, clasificar, explicar o generalizar. Si lo que se pretende medir es una «variable latente» o «constructo no observable», es necesario hacerlo de forma indirecta por medio de «indicadores observables» (p. ej., respuestas a preguntas de cuestionario o declaración de síntomas y percepciones). La bondad de esta medición depende de la relación entre estos indicadores observables y los constructos subyacentes. Si la relación es débil, las inferencias efectuadas serán imprecisas y probablemente incorrectas.

Este artículo se centra en los casos que requieran la cumplimentación de un cuestionario (por parte del paciente, el familiar o el personal médico) y la evaluación de éste como instrumento de medida. Se trata, pues, de la práctica habitual del clínico para determinar la presencia de una enfermedad mental en su paciente[1,2] o su grado de adherencia o cumplimiento del tratamiento[3]. También se trata de la práctica del investigador que quiere establecer la calidad de vida del hipertenso que sigue tratamiento con un nuevo fármaco[4] o del gerente del hospital que desea asegurarse de la calidad del servicio que ofrece[5]. Lo cual no es óbice para que esta metodología pueda emplearse también en la medición de variables fisiológicas como la presión arterial[6].

La creciente implicación de la opinión del propio paciente en las decisiones médicas ha creado la necesidad de medir los resultados de formas más ricas que la simple eficacia clínica y seguridad. Así, por ejemplo, la evaluación de la «calidad de vida relacionada con la salud», tal y como la perciben los pacientes, es una variable fundamental de lo que se conoce, en esta nueva óptica, como evaluación de resultados. La profusión de nuevos cuestionarios de calidad de vida, genéricos o específicos para determinadas afecciones, es una realidad corroborada por el número creciente de artículos publicados en revistas médicas generales y la existencia de diversas revistas especializadas.

Ante la proliferación de estudios para la validación de escalas relacionadas con la salud, se quiere ofrecer al lector algunas reflexiones desde la perspectiva de la psicometría (uno de cuyos ámbitos es el desarrollo y la validación de escalas) que le resulten útiles para poder juzgar la calidad de lo que, en ocasiones demasiado superficialmente, se califica como proceso de «validación» de un cuestionario. El objetivo de este artículo concierne únicamente a sus aspectos estadísticos. No se tratan las etapas previas, de carácter más conceptual, que determinarían tanto el contenido como el diseño de nuevos cuestionarios, así como su adaptación a otras culturas, sin las que toda sofisticación estadística es inútil. Véanse algunas de las guías existentes[7-12].

Propiedades fundamentales del instrumento de medida: validez y fiabilidad

Para validar un cuestionario coexisten esencialmente dos planteamientos estadísticos. Uno, enraizado en la teoría clá-

Correspondencia: Dr. J.M. Batista-Foguet.
ESADE. Dirección de Investigación. Universitat Ramon Llull.
Avda. Pedralbes, 60-62. 08034 Barcelona. España.
Correo electrónico: batista@esade.es

sica del test (TCT) y en modelos de análisis factorial, el otro, basado en la teoría de la respuesta al ítem (TRI). Aunque el uso de la TRI en investigación biomédica es creciente, su origen y desarrollo se deben a la medición de aptitudes o conocimientos. Por ello, este artículo presentará la actual perspectiva de la TCT, más cercana a la práctica médica.

Fiabilidad y validez proporcionan el lenguaje esencial de la medición y constituyen los índices de calidad de los cuestionarios. Ambas son cuestiones de grado. La fiabilidad tiene básicamente un cariz empírico y se centra en el rendimiento de las mediciones realizadas. Por el contrario, la validez tiene una orientación más teórica, pues inevitablemente emerge la cuestión: ¿para qué es válido? De hecho, no se valida un instrumento de medida en sí mismo, sino en relación con el propósito para el que se utilizará.

Validez

Entre los distintos tipos de validez (de aspecto; de contenido; de criterio –concurrente y predictiva–; y de constructo), esta última es la de mayor interés, ya que incorpora en gran parte las anteriores y es idónea para la evaluación de cuestionarios. A su vez, la validez de constructo se divide en nomológica, convergente y discriminante[13]. La validez nomológica se refiere a que las medidas válidas de diferentes conceptos teóricamente vinculados deben estar relacionadas de acuerdo con las teorías correspondientes. La validez convergente se refiere a que las medidas de un mismo concepto deben estar relacionadas, y deben estarlo más que las medidas de conceptos distintos, lo que constituye la validez discriminante.

La validez de constructo se evalúa habitualmente mediante correlaciones con otras escalas o indicadores objetivos de salud. En ocasiones, el análisis de la varianza, con mediciones pretratamiento y postratamiento, permite evaluar la sensibilidad al cambio o discriminar entre distintas poblaciones. En el mejor de los casos, un análisis factorial exploratorio (AFE) de la matriz de correlaciones «dictará» las dimensiones latentes y sus resultados se utilizarán como indicación de validez convergente y discriminante.

La calibración (en este mismo número)[14], habitual en la práctica médica dada su finalidad, podría asimilarse a la evaluación de la validez de criterio propia del cuestionario; es decir, el grado de acuerdo con una medida de referencia considerada más válida. No obstante, la imposibilidad de asumir medidas libres de error hace inviable esta estrategia cuando se evalúan cuestionarios.

Todo tipo de valoración de la validez supone, en esencia, preguntarse si los indicadores lo son sólo del concepto que se quiere medir y si no están influidos por ningún otro efecto sistemático. Sin embargo, las respuestas a los ítems de un cuestionario tienen un sinfín de efectos sistemáticos, de naturaleza diversa, que comprometen la validez[11]. La causa puede residir en el cuestionario, pero también en el encuestador, el encuestado o en el método de recogida de información. Puede tratarse de la modalidad de respuesta del cuestionario, de las expectativas del encuestador, de la reactividad frente a la situación, de la percepción de amenaza a la intimidad, de sucumbir al efecto de deseabilidad social o al de aquiescencia.

Fiabilidad

Mientras la invalidez se debe al error sistemático, la fiabilidad se relaciona con el grado de error aleatorio. Cuanto mayores son las fluctuaciones aleatorias en las respuestas, menor es la fiabilidad y viceversa. En la práctica, una medición es fiable cuando proporciona resultados consistentes o estables, ya sea en medidas repetidas o en las respuestas a los diversos ítems que la componen.

La medición de constructos data de principios del siglo XX[15,16] con la introducción, en principio entre psicólogos y posteriormente entre psiquiatras[17], de los métodos biométricos (regresión y correlación) de Galton y Pearson. Esta tradición estadística basada en la correlación, aunque se ha demostrado incorrecta en muchos casos[18], sigue utilizándose todavía en cuestionarios relacionados con la salud para estimar la fiabilidad y algunas formas de validez.

Así, se acostumbra entender la fiabilidad como la consistencia interna de los ítems que mide el coeficiente α de Chronbach (basado en el promedio de las correlaciones) o como la estabilidad temporal que proporcionan las correlaciones test-retest[12]. Carrasco y Jover discuten en este mismo número otras medidas usuales en medicina para evaluar la fiabilidad-concordancia[14].

Desde nuestro punto de vista, el binomio que constituyen el α de Chronbach y el AFE es del todo insuficiente para garantizar la validez y fiabilidad de un cuestionario relacionado con la salud. Por ello, este artículo, en primer lugar, critica la evaluación que habitualmente se hace de fiabilidad y validez, y en segundo lugar, propone el análisis factorial confirmatorio (AFC) –caso particular de los modelos de ecuaciones estructurales– como alternativa adecuada[19].

Métodos clásicos para evaluar la fiabilidad. Teoría clásica del test

Los psicólogos de finales del siglo XIX, preocupados únicamente por la falibilidad de las mediciones, desarrollaron la teoría clásica del test (TCT)[20]. Ésta establece la descomposición de la puntuación observada del j-ésimo ítem según la ecuación:

$$v_j = \lambda_{jk} f_k + e_j \qquad (1)$$

donde se asume que:

– f_k es «puntuación verdadera» y e_j error aleatorio de medición exclusivamente (sin componentes sistemáticos específicos del ítem v_j);
– las unidades de f_k y v_j son las mismas ($\lambda_j = 1$);
– f_k no está estandarizado y tiene varianza ϕ_{kk}.

lo que implica analizar la varianza de v_j en varianza explicada por la puntuación verdadera (ϕ_{kk}) y debida al error aleatorio θ_{jj}:

$$\sigma_j^2 = \phi_{kk} + \theta_{jj} \qquad (2)$$

La fiabilidad κ se define como el porcentaje de varianza de v_j explicado por f_k:

$$\kappa_j = 1 - \frac{\theta_{jj}}{\sigma_j^2} = \frac{\phi_{kk}}{\sigma_j^2} \qquad (3)$$

Según sus supuestos, de menos a más estrictos, los indicadores de una misma f_k pueden ser «congenéricos» (sin errores sistemáticos –derivados de cambios en la puntuación verdadera por funcionamiento distinto de los ítems o variaciones de opinión en el tiempo– ni errores correlacionados –por ejemplo, debido a efecto memoria–), «tau-equivalentes» (además con idénticas unidades, es decir, λ iguales) o «paralelos» (además con idéntica fiabilidad, es decir, θ iguales, y por consiguiente σ también iguales)[21].

En los siguientes apartados veremos que los investigadores que en medicina usan cuestionarios asumen implícitamente estos supuestos, al calcular la fiabilidad mediante la correlación de pares de ítems, o al utilizar el «coeficiente α de Chronbach»[22].

Cálculo de la fiabilidad de cada indicador-ítem

En efecto, conceptualizar la fiabilidad como estabilidad de las medidas (v_1, v_2) –repetición del mismo ítem en otro momento (*test-retest*), o bien administración de 2 ítems análogos simultáneamente– y calcularla, por tanto, como correlación entre éstas:

$$\rho_{v1v2} = \frac{\sigma_{v1v2}}{\sigma_{v1}\sigma_{v2}} = \frac{\lambda_{1k}\phi_{kk}\lambda_{2k}}{\sigma^2} = \frac{1\phi_{kk}1}{\sigma^2} \qquad (4)$$

supone, en realidad, que los ítems son paralelos, supuesto que puede etiquetarse eufemísticamente de poco realista[23].

Cálculo de la fiabilidad de escalas sumadas (conjunto de ítems)

Para reducir los efectos del error de medición del cuestionario se acostumbra sumar las puntuaciones de los ítems, de forma que los términos de error tiendan a compensarse en la «escala sumada» finalmente obtenida. Como es sabido, la fiabilidad de la escala es mayor que la de cada uno de los ítems que la componen, y tanto mayor cuanto mayor sea el número de ítems[24,25].

Para evaluar la fiabilidad de la escala, de forma análoga a como se ha hecho con el ítem individual, se consideran dos formas paralelas de un mismo test o bien su subdivisión en 2 mitades equivalentes *(split-halves),* y se correlacionan ambas a continuación. Dada la arbitrariedad del *split-halving* (la escala se puede partir de muchas maneras), los psicómetras han desarrollado coeficientes de fiabilidad que la evitan (aunque comparten sus supuestos) conocidos como coeficientes de consistencia interna o equivalencia. Entre ellos el más popular es el coeficiente α, para el que, de manera un tanto simplista, y sin hacer referencia a los supuestos mencionados, se proponen[12] umbrales mínimos como 0,7 para comparar grupos y 0,9 para comparaciones individuales. La diferencia entre estos umbrales está relacionada con la necesidad de disminuir la incertidumbre en el caso del diagnóstico de un solo paciente, mientras que la posibilidad de aumentar la precisión con un mayor tamaño muestral permite relajar el umbral en los estudios de grupos.

Crítica al modelo de la teoría clásica del test cuando evalúa la fiabilidad

Hemos visto que en medicina la fiabilidad se estima según correlaciones y asunciones poco justificadas en datos recogidos por encuesta, ya que éstos incluyen errores que no son aleatorios. Encuestador, encuestado y cuestionario son fuentes del error sistemático que conducen a vulnerar los supuestos de la TCT.

Además, estas asunciones tan restrictivas no se someten a ningún contraste. En particular, la medida de fiabilidad más utilizada (α de Chronbach) sólo estima correctamente la fiabilidad si los ítems son al menos tau-equivalentes[18,26] (cuyo contraste requiere comprobar la igualdad de covarianzas, no de correlaciones). En cualquier otro caso, proporciona un límite inferior de la fiabilidad. Sus defensores ponderan precisamente este aspecto conservador; no obstante, α se utiliza a menudo, en correlaciones o coeficientes de regresión, para corregir el sesgo debido a errores de medición, con lo que una fiabilidad subestimada se traducirá en una corrección excesiva del sesgo, lo que no es nada conservador.

Métodos clásicos para evaluar la validez. La perspectiva exploratoria del modelo de análisis factorial

Hasta finales de los años sesenta el investigador se ha servido del AFE[27] para establecer indicadores adecuados que hicieran emerger dimensiones subyacentes. Actualmente, todavía es frecuente en medicina (a pesar de disponer de mejores alternativas) servirse de estos modelos exploratorios para validar cuestionarios[28]. El AFE incluye un primer supuesto sustantivo: cada ítem tiene dos fuentes de variación, la «común» y la «única». Este supuesto se especifica mediante una ecuación de regresión, que relaciona los ítems (dependientes) con los factores, cuya naturaleza latente es precisamente lo que diferencia este modelo del de regresión. Estos m factores constituyen y explican la parte común o compartida por los ítems y se conocen como factores comunes. El término residual e_j es la parte única o sin explicar por los factores latentes:

$$v_j = \lambda_{j1}f_1 + \lambda_{j2}f_2 + \ldots + \lambda_{jm}f_m + e_j \qquad (5)$$

Esta expresión recuerda la (1), aquélla con un solo factor. Los parámetros λ_{jk} se denominan «saturaciones» y desempeñan un papel análogo a los coeficientes de regresión. La práctica habitual de estandarizar las saturaciones facilita su interpretación como correlación del ítem y el factor correspondiente.

Dado que, en general, el ítem no será un reflejo exacto de la información de los factores, el término residual e_j en la ecuación incluye dos tipos de efectos: los debidos a «características específicas del indicador», asociados a invalidez, y los del «error aleatorio de medición». Ambos efectos se asumen incorrelacionados entre sí y con los factores comunes. Estos supuestos permiten descomponer la varianza de cada ítem en «comunalidad» (varianza explicada por los factores comunes), y «unicidad» o varianza única, sin explicar por esos factores.

Crítica al modelo del AFE para evaluar la validez

Arrecian las críticas al modelo exploratorio y hasta estas páginas se han hecho eco de sus inconvenientes[29]. Aquí resaltamos las deficiencias que hemos denunciado en otros escritos:

– La propia formulación del AFE lo diferencia poco de otras técnicas multivariantes y descriptivas, llamadas genérica y equívocamente factoriales. De hecho, se identifica habitualmente la «técnica de análisis en componentes principales» (ACP) con un «modelo», el del AFE. Aquella técnica no asume ningún modelo para los ítems, su única finalidad es reducir la dimensionalidad.

– Los paquetes estadísticos, verdaderos promotores de este festival de la confusión, ni tan siquiera requieren que el usuario explicite el número de factores del modelo. Además, se acostumbra obviar su contraste estadístico y se utilizan en su lugar criterios sin fundamentar, propios del ACP, como umbrales mínimos para la varianza explicada[28].

– No es posible determinar una única expresión de cada ítem a partir de los factores comunes. Esta indeterminación propia del AFE se conoce con el eufemismo de «rotación»[30]. Cada método de rotación –por cierto, arbitrario– conduce a interpretaciones distintas y a menudo mantiene la rígida y poco realista estructura de factores incorrelacionados.

– El modelo incorpora pocos supuestos sustantivos y permite que cada ítem dependa de todos los factores comunes, con lo que la interpretación es heurística y difícil. En consecuencia, con el AFE es imposible demostrar la validez –que cada indicador mida únicamente el factor que se supone que debe medir.

Sin embargo, en investigación médica, la perspectiva exploratoria parece ser la única aproximación al replicar la estructura factorial de un cuestionario desarrollado en otro país. Aún hoy, desviarse de esta metodología tradicional en medicina es excepcional[31-33]. Y por si esto fuera poco, la secuencia de análisis procede evaluando en primer lugar la fiabilidad para después aplicar el AFE. De nada sirve una medida fiable de algo distinto a lo que se desea medir. El diagnóstico de la validez debe preceder siempre al de la fiabilidad. Además, basta con reunir un gran número de ítems inválidos, o incluso multidimensionales, para obtener un α[26] elevado.

Métodos alternativos para evaluar la validez y la fiabilidad. El análisis factorial confirmatorio

Cuando el investigador tiene suficientes conocimientos previos para formular hipótesis concretas sobre la relación entre indicadores y dimensiones latentes, su interés se centra en contrastar estas hipótesis. Por ejemplo, al traducir o adaptar cuestionarios ya desarrollados sabemos qué ítems deberían medir qué dimensiones. El modelo de análisis factorial confirmatorio (AFC)[34,18,19] corrige las deficiencias inherentes a la perspectiva exploratoria y conduce a una mayor concreción de las hipótesis que deben ser contrastadas. Su especificación difiere de la perspectiva exploratoria en aspectos esenciales como:

– Permitir restricciones en algunas saturaciones. Lo habitual es suponer la validez de cada ítem, es decir, que satura en un único factor. Se delimita así el concepto de factor común a aquel que subyace únicamente a sus indicadores concretos y se evita introducir factores *ad hoc* de difícil interpretación.

– Permitir contrastes estadísticos de las hipótesis especificadas.

– Permitir componentes únicas correlacionadas. Aunque es un recurso poco elegante, se justifica por la existencia de otros factores sin interés, como un método de medición común que no se desea explicitar en la especificación.

– Permitir analizar la matriz de covarianzas en lugar de la de correlaciones, indispensable para establecer si los indicadores son tau-equivalentes.

Adviértase que el AFC es mucho menos restrictivo que la TCT. En efecto, el AFC sólo asume que los ítems constituyen «mediciones congenéricas», pero no asume la igualdad de las saturaciones ni de las varianzas de error. Además, el AFC somete estos supuestos a contrastes estadísticos que, en caso de rechazarse, desaconsejarían la evaluación de la fiabilidad.

El modelo se suele representar en un diagrama de flujos *(path diagram),* acorde con su especificación. Convencionalmente, los rectángulos representan ítems y las elipses, factores comunes. Flechas unidireccionales entre factores comunes e ítems expresan saturaciones. Flechas bidireccionales indican correlaciones entre factores comunes o únicos. La figura 1 muestra los diagramas de dos posibles modelos de AFE y de AFC. En el modelo de AFC, los factores únicos de las variables v_1 y v_4 que podrían compartir método de medición están correlacionados. Se resalta que v_1, v_2 y v_3 son indicadores exclusivamente de f_1 mientras que v_4, v_5 y v_6 lo son sólo de f_2.

En un principio, los programas para estimar modelos de AFC eran escasos y requerían conocimientos de álgebra matricial. Actualmente, existe una gran variedad de ellos, todos accesibles y sencillos de utilizar (en algunos, el usuario se limita a dibujar el diagrama del modelo) que permiten estimar cualquier modelo de ecuaciones estructurales.

Evaluación de las propiedades del cuestionario mediante modelos de análisis factorial confirmatorio

El diseño ideal del cuestionario no consiste tanto en una única batería de ítems relativos al constructo global de interés como en subconjuntos de ítems específicos para cada dimensión.

La asignación de indicadores específicos a dimensiones concretas es una de las mayores aportaciones de la perspectiva confirmatoria. Los modelos de AFC permiten contrastar la validez ajustando un modelo que la asuma y diagnosticando su bondad de ajuste (validación de constructo). En este modelo, cada ítem satura únicamente sobre el factor-dimensión del que se supone que constituye un indicador válido. La invalidez de los ítems se detecta en indicios como los siguientes[19]:

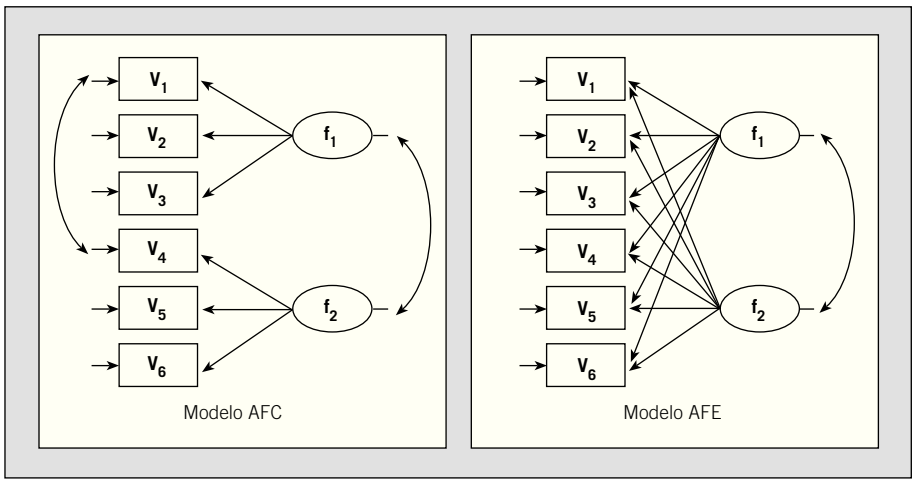

Fig. 1. Diferencias entre el diagrama de flujos (path diagram) *de los modelos de análisis factorial confirmatorio y análisis factorial exploratorio con 6 variables y 2 factores.*

– La bondad del ajuste del modelo es insatisfactoria, lo que implica que algunas saturaciones, o correlaciones entre factores únicos, se asumieron nulas por error. Los ítems que saturan en más de un factor no son indicadores válidos. Asimismo, la necesidad de una correlación entre factores únicos refleja errores sistemáticos y es otra señal de invalidez.
– Algunos ítems tienen un porcentaje de varianza única elevado, lo que hace poco creíble que dicha varianza contenga sólo error aleatorio de medición (invalidez convergente).
– Algunos factores presentan correlaciones muy próximas a la unidad, lo que plantea que estos factores representan una única dimensión (invalidez discriminante).
– Las correlaciones entre algunos factores son de signo o magnitud sorprendente según la teoría, lo que evidencia que las teorías disponibles son falsas o las variables miden factores distintos de los que se suponían (invalidez nomológica). Lamentablemente, se evalúan aquí simultáneamente las teorías y la validez. Se recomienda, por tanto, incluir en el modelo un cierto número de dimensiones potencialmente relacionadas con las de interés[35].

Una vez establecida la validez, puede procederse a evaluar la fiabilidad. Ésta puede calcularse simplemente como el porcentaje de varianza del ítem explicado por el factor. La fiabilidad de una escala sumada de los ítems de una misma dimensión puede calcularse según el coeficiente Ω[36]:

$$\Omega = 1 - \frac{\Sigma\theta_{jj}}{\text{Var (escala)}} \qquad (6)$$

donde el denominador es la varianza total de la escala y el numerador es la suma de las varianzas de error de todos los ítems de la escala.

Etapas en el ajuste de un modelo de análisis factorial confirmatorio

Esta sección refiere sucintamente las etapas a seguir para validar un cuestionario empleando el AFC. Debe tenerse en cuenta que el ajuste de modelos estructurales es un proceso complejo, del que sólo se destacan algunos de los aspectos más relevantes. El lector interesado puede acudir a las referencias citadas.

Especificación, identificación y estimación

Según se ha comentado, la especificación del modelo establece que cada variable satura sólo sobre el factor común que mide, que los factores comunes están correlacionados, y que los únicos están incorrelacionados.

Una vez especificado el modelo, se debe evaluar si es estimable. Esta etapa se conoce como de «identificación». En el caso del AFC, con carácter general, se requieren para cada factor al menos dos ítems que ni saturen en otro factor ni presenten componentes únicas correlacionadas. La precisión de las estimaciones mejora sustancialmente si se dispone de tres indicadores por factor. Dado que las propiedades de los estimadores son asintóticas, se recomiendan tamaños de muestra superiores a 200, aunque depende de las características del modelo[37].

Existe una multitud de procedimientos de estimación del modelo. Los métodos clásicos se basan en el criterio de la máxima verosimilitud, de acuerdo con el supuesto de normalidad multivariante de los ítems. Existen métodos alternativos para los ítems de nivel de medida ordinal[38,39] (como los de Likert) y contrastes robustos para el caso de los ítems no normales[40].

Diagnóstico de la bondad del ajuste

Un modelo correcto es aquel que sólo incorpora las restricciones y supuestos que se cumplen en la población, sin omisión de parámetros. Puesto que los modelos sobreparametrizados, que imponen pocas restricciones, suelen conducir a ajustes perfectos de los datos, un buen modelo implicará un compromiso entre la parquedad y la bondad del ajuste[41].

El diagnóstico de la bondad de ajuste es crucial para establecer la validez del cuestionario. En Batista y Coenders[42] puede encontrarse este desarrollo ampliado. La etapa de diagnóstico permitirá distinguir los modelos que ajusten flagrantemente mal los datos de aquellos modelos que los ajusten razonablemente bien, aunque de estos últimos puede haber muchos. La etapa de diagnóstico nunca será, pues, capaz de demostrar que un modelo es correcto, sino, a lo sumo, incapaz de demostrar que es incorrecto. En consecuencia, tampoco será capaz de demostrar que un cuestionario es válido, sino, incapaz de demostrar que es inválido.

El diagnóstico empieza por un examen general de la solución obtenida para detectar problemas graves como la presencia de estimaciones no admisibles o la falta de convergencia del algoritmo de estimación. A continuación, se establece de modo global la adecuación o no del modelo. Finalmente, se emplean diagnósticos detallados, parámetro a parámetro, para detectar partes del modelo cuya especificación no es idónea.

Un primer diagnóstico global del modelo es el «contraste de razón de verosimilitudes o estadístico χ^2». Su hipótesis nula establece que las restricciones del modelo son correctas. Su objetivo consiste en detectar posibles parámetros indebidamente omitidos en el modelo. Ya que un modelo no puede ser más que una aproximación a la realidad, la hipótesis que establece que el modelo es exactamente correcto será siempre falsa e incluso absurdo su contraste. Además, dado que los modelos de AFC suelen estimarse sobre muestras relativamente grandes, la potencia del contraste es a menudo elevada y conducirá a rechazar modelos por insignificantes errores de especificación. En la práctica interesará más cuantificar el grado de ajuste (o desajuste) del modelo que simplemente rechazar o no la hipótesis nula.

La lista de índices de bondad de ajuste es muy larga y queda fuera del alcance de este artículo[43]. Destacamos el «residuo estandarizado cuadrático medio (SRMR)», el «error cuadrático medio de aproximación (RMSEA)» y las medidas de bondad de ajuste basadas en el estadístico χ^2, reescalado de manera que tome valores entre 0 y 1. El más utilizado es el «índice de ajuste no normado (NNFI)» de Tucker y Lewis, que es independiente del tamaño muestral y tiene en cuenta la parquedad del modelo además de su bondad de ajuste[44]. Con la debida flexibilidad, el ajuste se considera aceptable si el SRMR y el RMSEA no alcanzan 0,05 y el NNFI supera 0,95.

Diagnóstico detallado, modificación del modelo y capitalización del azar

Difícilmente los modelos de AFC ajustan los datos en un primer contraste. Pero el diagnóstico no sólo permite evaluar el modelo, sino también sugiere maneras de mejorarlo. Así, la modificación del modelo[45] se ha convertido en práctica habitual para optimizar la bondad del ajuste al añadir parámetros, conseguir mayor parquedad eliminándolos o aumentar la validez mediante la supresión de ítems inapropiados («poda de ítems»).

El proceso de modificación viene guiado esencialmente por dos índices: «contraste de los multiplicadores de Lagrange»

(índice de modificación) y «contraste de Wald» (estadístico t). El primero refiere la significación de los parámetros omitidos del modelo. Éstos corresponden a saturaciones sobre otros factores o covarianzas entre factores únicos en el modelo de AFC y si fueran significativos indicarían invalidez. El segundo comprueba la significación de los parámetros incluidos en el modelo. Una saturación no significativamente distinta de 0 o una correlación entre factores no significativamente distinta de 1 indican invalidez.

Diversos autores[46,47] aconsejan introducir sólo modificaciones plausibles, de manera secuencial, reexaminando los resultados antes de efectuar la siguiente y empezar añadiendo parámetros significativos antes que eliminar los no significativos o los ítems poco válidos.

En cualquier caso, debe tenerse presente que la modificación del modelo se ha basado en los resultados de una muestra concreta. La introducción de modificaciones adecuadas para el ajuste del modelo a la muestra, pero inadecuadas para el ajuste a la población se denomina «capitalización del azar». La capitalización del azar tiende a sesgar al alza las estimaciones y los estadísticos t[48], lo que obliga a tener suma cautela en la interpretación. Para que los resultados y el modelo final puedan generalizarse más allá de la muestra concreta, precisaremos estimar y diagnosticar el modelo en una segunda muestra independiente («validación cruzada»).

Discusión

El cuestionario es una pieza cada vez más relevante en la evaluación de la evidencia científica a favor de los efectos de un tratamiento, de la satisfacción del usuario o de las preferencias por diversos resultados sanitarios. Por ello, el investigador en medicina debe conocer qué procedimientos estadísticos le permiten hoy optimizar la medida de sus variables de interés.

Es obvio que haber probado un cuestionario en una muestra de pacientes, haber calculado algunas correlaciones y publicado sus resultados, no garantiza que éste se haya validado. El proceso de validación de un cuestionario implica un conjunto de decisiones que se apoyan en contrastes de hipótesis correctamente formuladas. Además, cada cuestionario requiere de la suficiente evidencia científica para su adecuación.

Las publicaciones biomédicas que incluyen cuestionarios revelan la necesidad de precisar, como hemos hecho, los términos esenciales de la medición (fiabilidad y validez). Como se ha visto, la investigación con cuestionarios o tests utiliza generalmente metodologías tradicionales basadas en supuestos poco realistas que se incumplen en la práctica y, en consecuencia, la fiabilidad se evalúa incorrectamente y, habitualmente, antes de que se evalúe la validez.

Se han señalado las limitaciones de los procedimientos de evaluación de un cuestionario mediante la TCT. En contraposición, se han referido las ventajas de los modelos del AFC, que permiten evaluar la validez y la fiabilidad de cada ítem. El investigador contrasta hipótesis acerca de ítems individuales, ¿miden lo que pretenden medir? (validez) y, una vez establecida ésta, ¿con qué precisión se obtiene esta medida? (fiabilidad). Los modelos del AFC han mostrado su aptitud tanto para validar un nuevo cuestionario como para adaptar el desarrollado en otra lengua o aplicado a una población o cultura diferente, puesto que nada garantiza que los ítems se entiendan de la misma manera en distintos contextos, o incluso que las variables latentes tengan la misma conceptualización[8,12].

REFERENCIAS BIBLIOGRÁFICAS

1. Goldberg DP. The detection of psychiatric illness by questionnaire. London: Oxford University Press, 1972 (Maudsley Monograph n.º 21).
2. Muñoz PE, Vázquez JL, Pastrana E, Rodríguez F, Oneca C. Study of the validity of Goldberg's. Soc Psychiatry 1978;13:99-104.
3. Knobel H, Alonso J, Casado JL, Collazos J, González J, Ruiz I, et al. Validation of a simplified medication adherence questionnaire in a large cohort of HIV-infected patients: the GEEMA study. AIDS 2002;16:605-13.
4. Croog SH, Levine S, Testa MA, Brown B, Bulpitt CJ, Jenkins CD, et al. The effects of antihypertensive therapy on the quality of life. NEJM 1986;314:1657-64.
5. Murphy J, Chang H, Montgomery JE, Rogers WH, Safran DG. The quality of physician-patient relationships. Patients' experiences 1996-1999. J Fam Parct 2001;50:123-9.
6. Batista-Foguet JM, Coenders G, Artés M. Using structural equation models to evaluate the magnitude of measurement error in blood pressure. Statistics in Medicine 2001;20:2351-68.
7. Behling O, Law KS. Translating questionnaires and other research instruments: problems and solutions. Thousand Oaks: SAGE, 2000.
8. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of healthrelated quality of life measures: literature review and proposed guidelines. J Clin Epidemiol 1993;46:1417-32.
9. Hunt SM, Alonso J, Bucquet D, Niero M, Wiklund I, McKenna S. Cross-cultural adaptation of health measures. Health Policy 1991;19:33-44.
10. Sudman S. Thinking about answers: the application of cognitive processes to survey methodology. San Francisco: Jossey-Bass, 1996.
11. Groves RM. Survey errors and survey costs. New York: John Wiley & Sons, 1989.
12. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: Attributes and review criteria. Qual Life Res 2002;11:193-205.
13. Campbell DT, Russo MJ. Social Measurement. Thousand Oaks: SAGE, 2001.
14. Carrasco JL, Jover L. Métodos estadísticos para evaluar la concordancia. Med Clin (Barc) 2003;121(supl. 2):28-34..
15. Spearman C. General intelligence, objectively determined and measured. Am J of Psychol 1904;15:201-93.
16. Brown W. The essentials of mental measurement. Cambridge: Cambridge University Press, 1911.
17. Spearman C. The tenth Maudsley Lecture: The psychiatric use of methods and results of experimental psychology. An investigation into the significance of perseveration. Journal of Mental Sciences 1928;34:653-9.
18. Bollen KA. Structural equations with latent variables. New York: John Wiley & Sons, 1989.
19. Batista-Foguet JM, Coenders G. Introducción a los modelos estructurales. Utilización del análisis factorial confirmatorio para la depuración de un cuestionario. En: Renom J, editor. Tratamiento informatizado de datos. Barcelona: Masson, 1998; p. 229-86.
20. Nunnally JC. Psychometric theory. 2nd ed. New York: McGraw-Hill, 1978.
21. Jöreskog KG. Statistical analysis of sets of congeneric tests. Psychometrika 1971;36:109-33.
22. Chronbach LJ. Coefficient alpha and the internal structure of the test. Psychometrika 1951;16:297-334.
23. Coenders G, Saris WE, Batista-Foguet JM, Andreenkova A. Stability of three-wave simplex estimates of reliability. Structural Equation Modelling 1999;6:135-57.
24. Spector PE. Summated rating scale construction. 1st ed. Thousand Oaks: Sage, 1992.
25. Batista-Foguet JM, Fortiana J, Currie C, Villalbi JR. Socio-economic indexes in surveys for comparisons between countries. An applied comparison using the family affluence scale. Social Indicators Research [en prensa].
26. Cortina JM. What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology 1993;78:98-104.
27. Lawley DN, Maxwell AE. Factor analysis as a statistical method. 1st ed.London: Butterworth, 1971.
28. Macías Robles MD, Fernández-López JA, Hernández-Mejía R, Cueto-Espinar A, Rancaño I, Siegrist J. Evaluación del estrés laboral en trabajadores de un hospital público español. Estudio de las propiedades psicométricas de la versión española del modelo «desequilibrio esfuerzo-recompensa». Med Clin 2003;120:652-7.
29. Prieto L, Alonso JA. Propósito del uso del análisis factorial en la evaluación de la equivalencia transcultural de cuestionarios. Med Clin 1998;19:717-8.
30. Batista-Foguet JM. Análisis en componentes principales y modelo de análisis factorial. En: Sánchez Carrión, editor. Métodos de análisis multivariantes aplicados en ciencias sociales. Madrid: CIS-Siglo XXI, 1984.
31. Keller SD, Ware JE, Bentler PM, Aaronson NK, Alonso J, Apolone G, et al. Use of structural equation modeling to test the construct validity of the SF-36 Health Survey in ten countries: results from the IQOLA project. J Clin Epidemiol 1998;51:1179-88.
32. Bentler PM, Stein JA. Structural equation models in medical research. Statistical Methods in Medical Research 1992;1:159-81.
33. Dunn, G. Statistics in Psychiatry. London: Arnold, 2000.
34. Jöreskog KG. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika 1969;34:183-202.
35. Chronbach LJ. Mehl PE. Construct validity in psychological tests. Psychological Bulletin 1955;52:281-302.

36. Heise DR, Bohrnstedt GW. Validity, invalidity and reliability. En: Borgatta EF y Bohrnstedt GW, editors. Sociological methodology 1970. San Francisco: Jossey-Bass, 1970; p. 104-29.
37. Boomsma A, Hoogland JJ. The robustness of LISREL modeling revisited. En: Cudeck R, Du Toit S, Sörbom D, editors. Structural equation modeling: present and future. A festschrift in honor of Karl Jöreskog. Chicago: Scientific Software International 2001; p. 139-68.
38. Jöreskog KG. New developments in LISREL. Analysis of ordinal variables using polychoric correlations and weighted least squares. Quality & Quantity 1990;24:387-404.
39. Coenders G, Satorra A, Saris WE. Alternative approaches to structural modelling of ordinal data. A Monte Carlo study. Structural Equation Modelling 1997;4:261-82.
40. Satorra A, Bentler PM. Scaling corrections for chi-square statistics in covariance structure analysis. En: Van Eye A, Clogg CC, editors. Latent variable analysis. Thousand Oaks: SAGE, 1994; p. 399-419.
41. Box GEP. Science and statistics. Journal of the American Statistical Association 1976;71.791-9.
42. Batista-Foguet JM, Coenders G. Modelos de ecuaciones estructurales. Modelos para el análisis de relaciones causales. 1.ª ed. Madrid: La Muralla, 2000.
43. Bollen KA, Long JS. Testing structural equation models. 1.ª ed. Thousand Oaks: SAGE, 1993.
44. Marsh HW, Balla JR, Hau KT. An evaluation of incremental fit indices. A clarification of mathematical and empirical properties. En: Marcoulides GA, Schumacker RE, editors. Advanced structural equation modeling. Issues and techniques. Mahwah. New Jersey: Lawrence Erlbaum, 1996; p. 315-53.
45. Muñoz J. Análisis factorial confirmatorio y capitalización del azar. Una aplicación práctica. Papers ESADE 1999;14:1-33.
46. MacCallum RC, Roznowski M, Necowitz LB. Model modification in covariance structure analysis: the problem of capitalization on chance. Psychological Bulletin 1992;111:490-504.
47. Saris WE, Stronkhorst LH. Causal modelling in nonexperimental research. Amsterdam: Sociometric Research Foundation, 1984.
48. Luijben T. Statistical guidance for model modification in covariance structure analysis. Amsterdam: Sociometric Research Foundation, 1989.