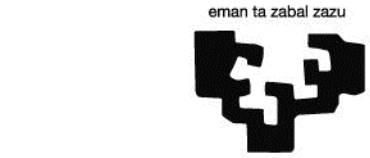


Validación y reducción de cuestionarios de calidad de vida relacionada con la salud: aplicación de diferentes metodologías

Amaia Bilbao González

2016

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Validación y reducción de cuestionarios de calidad de vida relacionada con la salud: aplicación de diferentes metodologías

Amaia Bilbao González

2016

Tesis doctoral realizada bajo la dirección de:

Dra. Inmaculada Arostegui Madariaga y

Dr. José María Quintana López

A mis aitas, Asier y Loli, por todo lo que me habéis enseñado a lo largo de mi vida, y lo que aún me seguís enseñando. Por todos los valores que me habéis inculcado. Por haberme transmitido la ilusión por trabajar y aprender. Por vuestra plena confianza en mí. Por vuestra infinita paciencia y comprensión. Por ser como sois. Os quiero muchísimo.

A mis hermanos, Itziar y Jabi, por estar siempre ahí, preocupados y pendientes de mi. A ti Itzi, que decirte. No sé que haríamos la una sin la otra. Eres mi hermana, mi mejor amiga, y probablemente la persona que mejor me conoce. Con sólo mirarme ya sabes lo que me pasa. Gracias por estar siempre a mi lado, por escucharme cada día, por entenderme tan bien, y por todo el amor y cariño que me transmites.

A mis chicos, Exipion y Aimar. A ti Exipion, por todo lo que hemos vivido y aprendido juntos, que no ha sido poco. Por toda tu paciencia a lo largo del camino, por todo el apoyo que siempre me has mostrado, y por tu preocupación diaria. Por haberme respetado periodos de ausencia dedicados a la investigación y no a ti. Por todo lo que valoras mi trabajo y lo orgulloso que te sientes de mí. Y por la ilusión que te hace que sea Doctora. A mi txikitxu Aimar, por tu sonrisa, cariños y musutxus diarios. Por todo el tiempo que no te he podido dedicar. Aunque aún no te des cuenta, estoy segura de que algún día te sentirás orgulloso de tu amatxu. A los dos, gracias de todo corazón.

A todos, os quiero muchísimo.

Agradecimientos

A Txema Quintana e Inma Arostegui, directores de esta tesis, quiero agradeceros profundamente toda la confianza que habéis puesto siempre en mí. Gracias por vuestro apoyo incondicional, no solo en el desarrollo de esta tesis, sino en el desarrollo de toda mi carrera profesional. Con vosotros comencé hace años mi carrera investigadora, y vosotros fuisteis los que me inculcasteis la inquietud por la investigación. Siempre me habéis guiado en este camino. Os agradezco vuestra dedicación, escucha, paciencia, vuestros sabios consejos, vuestro constante estímulo para que siga adelante, y el cariño que siempre me habéis mostrado. Todo ello ha hecho posible que haya terminado esta tesis. He aprendido mucho de vosotros. Gracias de todo corazón.

A los miembros de los equipos investigadores de los proyectos que han tomado parte en esta tesis, por vuestro trabajo, dedicación y colaboración mostrada en todo momento. Sin el trabajo de todos ellos esto no hubiera sido posible. Y en especial, a ti Antonio Escobar, mi jefe, por tu paciencia, buenos consejos y escucha

siempre que lo he necesitado. Gracias por la acogida que siempre me has dado en esta Unidad de Investigación, que ha hecho que me sienta tan a gusto.

A todo el equipo de la Unidad de Investigación del Hospital de Galdakao-Usansolo, donde comenzó todo este trabajo. Aunque hace años que ya no trabajo allí, siempre me acogéis como una más. Y en especial, quiero agradecer a Nerea González y Susana García, por todo el apoyo que siempre me habéis dado, por haberme escuchado siempre que lo he necesitado, por los buenos momentos que me habéis hecho pasar, y sobre todo por vuestra amistad. También quiero agradecer a Carlota Las Hayas, que aunque ya no trabaja en esta Unidad de Investigación, fue compañera de fatigas en gran parte del trabajo de esta tesis. Tengo que agradecer tu ayuda, tu paciencia y tu disponibilidad siempre que te he necesitado. Gracias por hacerme reír tanto. Ha sido un placer trabajar contigo y sobre todo haberte conocido.

A mis actuales compañeros de la Unidad de Investigación del Hospital Universitario Basurto, por vuestra cálida acogida. Me he sentido una más desde que entré. También os quiero agradecer vuestra paciencia y escucha, sobre todo en mis momentos de desespero.

A María Pertika, mi antigua compañera de BIOEF. Todavía recuerdo nuestros comienzos en la Fundación. Quiero agradecerte todo el apoyo moral que siempre he recibido de ti a lo largo de mi carrera. Aunque trabajamos en campos completamente distintos, siempre me has entendido. Gracias por escucharme, y sobre todo, gracias por tu amistad. También quiero agradecer al resto de mis antiguos compañeros de BIOEF por el apoyo que siempre me mostraron, ya que gran parte de este trabajo fue desarrollado durante aquel periodo. En especial, quiero agradecer a mi antigua jefa Carmen Garaizar, que siempre me animó a seguir adelante con la tesis. Sé que te hará ilusión.

A todos mis amigos, por ser como sois, por lo buenos momentos que me habéis hecho pasar, por el apoyo y cariño que siempre me habéis mostrado, por vuestra

paciencia, y por haberme escuchado a pesar de no saber muy bien lo que he estado haciendo a lo largo de todo este camino.

Quiero agradecer a los Servicios de Traumatología y Cirugía Ortopédica, los Servicios de Psiquiatría, así como los Servicios de Cirugía General y Aparato Digestivo de Osakidetza que han participado en los proyectos de investigación incluidos en esta tesis, por su colaboración, dedicación y esfuerzo en el desarrollo de esta investigación. También quiero agradecer a las personas que se encargaron de hacer la recogida de datos, y por supuesto, a los pacientes que han contestado a los cuestionarios, sin cuya colaboración este trabajo no se habría llevado a cabo.

Por último, quiero agradecer al Instituto de Salud Carlos III, al antiguo Departamento de Educación, Universidades e Investigación del Gobierno Vasco, al Departamento de Salud del Gobierno Vasco, y al Fondo Europeo de Desarrollo Regional (FEDER), por la contribución económica prestada para poder llevar a cabo estos proyectos de investigación.

Índice

Resumen

En las últimas décadas, la medición de la calidad de vida relacionada con la salud (CVRS) ha ido convirtiéndose en una disciplina autónoma y formal, de gran utilidad dentro del ámbito sanitario, y más concretamente dentro de las enfermedades crónicas. La manera de medir la CVRS es a través de cuestionarios y para establecer la aceptabilidad de un cuestionario es necesario que cumpla con ciertos atributos, entre los que se encuentran las propiedades psicométricas, tales como la fiabilidad, la validez y la sensibilidad al cambio. De entre las diferentes propiedades psicométricas, la validez de constructo, que consiste en proporcionar evidencia de que la forma de interpretar las puntuaciones del instrumento es correcta según la teoría y los constructos que se miden, es el aspecto más susceptible de exploración mediante el análisis numérico y estadístico. Este, a su vez, se divide en la validez convergente y discriminante, la validez de grupos conocidos y la validez de estructura, siendo esta última la que juega un papel más importante, y siendo también el tema con mayor controversia. Existen diferentes perspectivas y teorías de enfoque, desde la teoría clásica del test (TCT) hasta la teoría de la respuesta al ítem (TRI), e incluso dentro de cada una de las teorías, existen diferentes métodos. Sin embargo no hay un consenso sobre que

metodología utilizar para evaluar la validez de estructura. Por otro lado, aunque existen cuestionarios de CVRS con buenas propiedades psicométricas, a menudo su longitud limita en gran medida su aplicación, por lo que disponer de una versión reducida que mantenga las mismas buenas propiedades psicométricas que su versión larga, es de gran utilidad tanto en la practica clínica como en investigación. Al igual que ocurre en el proceso de validación de instrumentos de CVRS, en el proceso de reducción de cuestionarios los métodos más utilizados han sido métodos de la TCT. Sin embargo, la literatura más actual también propone para la reducción de cuestionarios la utilización de procedimientos más modernos basados en la TRI. Así, los objetivos generales de esta tesis son: 1) validar y/o reducir cuestionarios de CVRS mediante la combinación de métodos estadísticos de la TCT y de la TRI para evaluar así la complementariedad de dichos métodos; y 2) proporcionar herramientas de medición de la CVRS científicamente validadas y con buenas propiedades psicométricas en castellano para diferentes ámbitos de la salud.

Para llevar a cabo este trabajo, hemos seleccionado tres cuestionarios de CVRS para tres enfermedades crónicas diferentes: obesidad mórbida, artrosis de cadera y trastorno de la conducta de la alimentación (TCA). Los cuestionarios empleados fueron el cuestionario Obesity-related Problems scale (OP) diseñado para medir el impacto de la obesidad mórbida en el funcionamiento psicosocial, que consta de ocho preguntas cubriendo un único dominio; una versión reducida del cuestionario Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), empleado para la medición de sintomatología y función en pacientes con artrosis de extremidad inferior, que consta de 11 ítems que se conforman en dos factores, dolor y función; y el cuestionario Health Related Quality of Life for Eating Disorders version 2 (HeRQoLEDv2), empleado para medir el impacto del TCA sobre aspectos físicos, mentales y sociales, que consta de 55 preguntas que se conforman en nueve factores.

Para evaluar la fiabilidad se estudió la consistencia interna de los diferentes instrumentos, que hace referencia a la homogeneidad de los ítems que componen

una escala. El estudio de la validez convergente y discriminante de las escalas se llevó a cabo mediante la correlación de las escalas del instrumento con las escalas de otros instrumentos, estableciendo hipótesis previas en relación a que escalas del instrumento a validar deberían correlacionar alto o bajo con que escalas de otros instrumentos (validez convergente y discriminante, respectivamente). El estudio de validez de grupos conocidos se llevó a cabo comparando las puntuaciones de las escalas entre grupos de pacientes en los que se anticipaba que hubiera diferencias. La validez de estructura, que trata de realizar un análisis interno del instrumento para comprobar que la estructura hipotética subyacente del instrumento es correcta y adecuada, se llevó a cabo mediante métodos de la TCT y de la TRI. Los métodos de la TCT son modelos factoriales, que proporcionan combinaciones lineales de los ítems con los factores latentes que han de ser estimados mediante sumas ponderadas que reflejen la importancia de cada ítem en el constructo. Las técnicas que hemos empleado son el Análisis Factorial Exploratorio (AFE), el Análisis Factorial Confirmatorio (AFC), los Modelos de Ecuaciones Estructurales (MEE), así como el caso específico de estos modelos para datos categóricos. Aunque los tres casos, AFE, AFC, MEE, sirven para evaluar la dimensionalidad (número de factores) que es necesaria para representar la variabilidad que hay en los datos, el AFC y MEE permiten además contrastar estos patrones para confirmar la validez de los constructos hipotetizados, mediante diferentes índices de bondad de ajuste. Los métodos de la TRI son modelos probabilísticos que se basan en que un individuo con un nivel concreto de calidad de vida tendrá una cierta probabilidad de responder de forma positiva a una pregunta. Por tanto, la hipótesis subyacente en la TRI es que la respuesta a un ítem de un cuestionario depende del nivel de dificultad que el ítem presente y el nivel de habilidad respecto al constructo del individuo que responde. Estos modelos incluyen un conjunto de asunciones entre las que se encuentra la unidimensionalidad, ya que asume que los ítems de la escala son de una única dimensión o constructo. Entre los modelos de la TRI, hemos utilizado métodos para ítems de respuesta politómica, y más concretamente el Rating Scale Model y Graded Response Model. El primero se clasifica dentro de los modelos de 1–parámetro, ya que el único parámetro que caracteriza al ítem es su nivel de

dificultad. Y el segundo se clasifica dentro de los modelos de 2–parámetros, ya que además de la dificultad, también tienen en cuenta la discriminación del ítem. Por último, para evaluar la sensibilidad al cambio se compararon las puntuaciones basales y las de seguimiento, además de utilizar los tamaños del efecto para medir la magnitud del cambio.

En relación a los pacientes con obesidad mórbida, se reclutaron 123 pacientes que se encontraban en lista de espera para ser intervenidos de cirugía bariátrica, y se realizó la traducción, adaptación y validación del cuestionario OP al español, estudiando su fiabilidad y validez mediante la aplicación de diferentes métodos de la TCT. Los resultados demostraron la fiabilidad, la validez convergente y la validez de grupos conocidos. La validez de estructura se analizó mediante el AFE y el AFC, obteniendo resultados muy similares, además de índices de bondad de ajuste del AFC que confirmaron la estructura subyacente del cuestionario.

En la evaluación de las propiedades psicométricas del WOMAC, se seleccionaron dos cohortes independientes de pacientes (788 y 445 pacientes, respectivamente) con artrosis de cadera que estaban en lista de espera para ser intervenidos de artroplastia total de cadera, a los que además se les hizo un seguimiento a los seis meses después de la intervención. Se realizó la validación de la versión reducida propuesta del cuestionario WOMAC, estudiando la fiabilidad, validez y sensibilidad al cambio, combinando métodos de la TCT así como de la TRI en ambas cohortes de pacientes. Los resultados apoyaron la fiabilidad del instrumento, la validez convergente y la validez de grupos conocidos. Para el estudio de la validez de estructura se utilizó el AFC, que confirmó la existencia de dos factores en el cuestionario, dolor y función. Además, se aplicó el Rating Scale Model de la TRI a cada una de las dos escalas, concluyendo que ambas escalas se ajustaban al modelo Rasch, confirmando así la unidimensionalidad. Los resultados del AFC y del Rating Scale Model fueron muy similares en ambas cohortes de pacientes, demostrando así la estabilidad de los resultados. Por último, los parámetros de sensibilidad al cambio demostraron grandes cambios a los seis meses de la intervención.

En relación a los pacientes con un TCA, en primer lugar se reclutó una muestra de 324 pacientes, a través de la cual se confirmó la existencia de una estructura interna de segundo orden con dos medidas resumen en el cuestionario HeRQoLEDv2, para así posteriormente reducir dicho cuestionario. Se hipotetizó que 40 de los ítems del cuestionario que se conformaban en siete factores denominados factores de primer orden, a su vez se resumían en dos factores de segundo orden. Se empleó un MEE, cuyos resultados confirmaron la multidimensionalidad de dicho cuestionario, garantizando la existencia de dicha estructura de segundo orden con dos medidas resumen denominadas "Adaptabilidad social" y "Salud Mental y rendimiento". Se aplicó el Rating Scale Model a cada una de las escalas de forma separada para reducir cada una de las dos medidas resumen, garantizando la unidimensionalidad, dando lugar a la versión reducida de 20 ítems, denominada Health Related Quality of Life for Eating Disorders Questionnaire-Short Version (HeRQoLED-S). Posteriormente, se reclutó otra muestra de 377 pacientes con un TCA, a la que se realizó un seguimiento de un año, a través de la cual se realizó la validación de la versión reducida HeRQoLED-S, combinando métodos de la TCT así como de la TRI. Por un lado, se utilizó los MEE para datos categóricos, cuyos resultados confirmaron la estructura de segundo orden del cuestionario HeRQoLED-S. Por otro lado, se aplicó el Graded Response Model a cada una de las dos medidas resumen, cuyos resultados demostraron la unidimensionalidad de cada escala y la correcta capacidad de discriminación de todos los ítems de la escala con respecto a su correspondiente constructo latente. La validez convergente, la validez de grupos conocidos, la fiabilidad y la sensibilidad al cambio del instrumento reducido también quedaron demostradas.

En vista a los resultados obtenidos podemos concluir, en primer lugar, que los métodos basados en la TCT y los basados en la TRI son complementarios entre sí. Mientras que los primeros nos proporcionan información sobre la dimensionalidad del cuestionario, los segundos se centran en la unidimensionalidad. Así, cada uno de los métodos proporciona diferente tipo de soporte en el proceso de validación y/o reducción de cuestionarios de CVRS, de

forma que la combinación de los diferentes métodos proporciona mayor evidencia de la validez de cuestionarios de CVRS. En segundo lugar, podemos concluir que la utilización combinada de estas técnicas estadísticas ha proporcionado versiones científicamente validadas en castellano de tres cuestionarios de CVRS, para diferentes ámbitos de la salud, que pueden ser utilizados en la práctica clínica, así como entre investigadores y decisores sanitarios. Concretamente, la versión española del cuestionario OP posee buenas propiedades psicométricas de validez y fiabilidad, y puede utilizarse para medir el impacto de la obesidad mórbida en el funcionamiento psicosocial en estudios transversales. La versión reducida propuesta del cuestionario WOMAC es fiable, valida y sensible al cambio, y puede ser utilizada para medir dolor y función en pacientes con artrosis de cadera. La versión reducida HeRQoLED-S, posee dos medidas resumen, y es fiable, valida y sensible al cambio, proporcionando así una herramienta simple, corta y de fácil aplicación para medir la CVRS en pacientes con un TCA.

Capítulo 1

Introducción, antecedentes y justificación

1.1. Introducción

Inicialmente el concepto de la salud se definía como la ausencia de la enfermedad, pero ya en el año 1948, la Organización Mundial de la Salud consideró que esta definición era demasiado limitada, y propuso que la salud engloba muchos más aspectos. A partir de aquel momento, se definió la salud como "un estado de completo bienestar físico, mental y social, y no meramente la ausencia de enfermedad o discapacidad" (World Health Organization, 1952). Es de esta manera que se introdujo dentro del marco de la salud la concepción de la calidad de vida como la experiencia subjetiva del individuo de su bienestar físico, mental y social. El modelo "biosicosocial" en medicina enmarca aspectos referidos al bienestar del paciente, como sus relaciones como persona, su comportamiento, el entorno en el que se desenvuelve y sus relaciones sociales, en lo que se conoce con el nombre de calidad de vida (Sanz, 1991). Aunque el concepto de calidad de vida no es nuevo,

en la década de los 80 se produjo un interés popular y médico sobre la calidad de vida, que ha ido creciendo a lo largo de estos años. De esta forma, la calidad de vida, cuya característica básica es la subjetividad, ya que cada individuo es el único medidor de su percepción de la calidad de vida, basada en el sentido de bienestar de cada persona derivado de la experiencia diaria de su vida, se abre camino entre la medicina y las ciencias sociales.

La calidad de vida, considerada globalmente, es difícil de definir, ya que depende en gran medida de la escala de valores por la que cada individuo ha optado, y de los recursos emocionales y personales de cada uno. Además, está sometida a determinantes económicos, sociales y culturales y se modifica, con el paso del tiempo, para un mismo individuo. Por otro lado, cuando hablamos de calidad de vida, no solo hablamos de la calidad de vida en general, sino también de síntomas y efectos secundarios que pueden o no afectar en la calidad de vida. De esta manera, para distinguir la calidad de vida en su sentido más general de su significado en medicina clínica, y con el fin de evitar ambigüedades, comienza a desarrollarse un nuevo concepto, el de calidad de vida relacionada con la salud (CVRS) o salud percibida (Fayers y Machin, 2007). Revicki y cols. (2000) definen la CVRS como "la evaluación subjetiva que hace la persona del impacto que su enfermedad y su tratamiento tienen en los dominios físicos, psicológicos y sociales del funcionamiento y el bienestar". Así, la medición de la CVRS hace posible obtener información sobre la enfermedad y su impacto en la vida del paciente (Goldsmith, 1972 y 1973). Aunque algunos investigadores prefieren hacer énfasis solo en aspectos de la salud, definiendo así la CVRS, otros están adoptando el término de resultados percibidos por los pacientes (Patient reported outcomes), ya que este término incluye toda una serie de resultados, desde resultados como el dolor, fatiga, y depresión, hasta síntomas físicos como nauseas y vómitos (Fayers y Machin, 2007).

En los últimos 30 años la medición de los estados de salud y de la CVRS ha ido cogiendo cada vez mas importancia, y ha ido convirtiéndose en una disciplina autónoma y formal, de gran utilidad dentro del ámbito tanto sanitario como

político (Aaronson y cols., 2002). Así, en 1992 se creó la Fundación "Medical Outcomes Trust" (MOT), formada por organizaciones sin ánimo de lucro, investigadores académicos, agencias del sector público y firmas comerciales con el fin de promocionar la ciencia, y promocionar la aplicación de la medición de los resultados percibidos por los pacientes. El estado de salud de un individuo es el resultado global de la unión de dos componentes básicos, los componentes de salud clásicos de los que se encarga la medicina, y los componentes psicosociales que se evalúan conocido el punto de vista del paciente (Koller y Lorenz, 2002). Estos dos componentes han de operativizarse de tal manera que puedan ser comprobados empíricamente y sean relevantes clínicamente. Por tanto, para conocer la imagen completa del estado del individuo será necesario conocer su estado en ambos componentes.

La medición de la CVRS puede ser útil como una medida resultado. La medición del cambio producido en la CVRS puede ser útil como resultado de la terapia dada a un paciente. En el contexto de la salud y el manejo de la enfermedad, la medición de la CVRS puede ser de utilidad para promover la participación de los pacientes en el proceso de toma de decisiones de tratamiento, mejorando así la efectividad en la comunicación entre médico y paciente, y aumentando la satisfacción del paciente con el cuidado recibido. Los médicos pueden utilizar herramientas de medida de CVRS para controlar la progresión de la enfermedad de los pacientes e identificar problemas inesperados (especialmente del funcionamiento psicosocial) y otras cuestiones, como consecuencia directa o indirecta, de una condición o tratamiento (Huang y Speight, 2013). Medir la CVRS también puede resultar de utilidad a nivel de gestión política, para financiar un tratamiento en menos cabo de otro, aduciendo que ese tipo de tratamiento produce un mayor aumento de la CVRS, que el otro. Otras utilidades de la evaluación de la CVRS están relacionadas con la asignación de recursos económicos. Estos instrumentos proporcionan evidencia acerca de la efectividad de los tratamientos desde el punto de vista del paciente. Así, cada vez mas responsables de la salud atienden a este tipo de mediciones para decidir donde destinar estos fondos monetarios (Bowling, 2001). Finalmente, dentro de la industria farmacéutica, la medición de la CVRS es cada vez más

popular, ya que la emplean en sus ensayos clínicos con nuevos fármacos. Si la CVRS del individuo ha mejorado tras la administración del fármaco, este prueba tener un buen efecto en la vida del individuo (Huang y Speight, 2013).

Esta tesis doctoral trata sobre la validación y reducción de cuestionarios de CVRS y se presenta en nueve capítulos. En el resto de este primer capítulo, inicialmente se introducen los cuestionarios de CVRS, así como los atributos y propiedades psicométricas que estos cuestionarios han de cumplir, tales como la fiabilidad, validez y sensibilidad al cambio. Posteriormente, se describen los diferentes métodos para la evaluación de la validez y/o reducción de instrumentos de CVRS, tanto métodos de la psicometría clásica como de la moderna. Finalmente, se introduce la medición de la CVRS en enfermedades crónicas de diferentes ámbitos sanitarios, a saber patología psiquiátrica, osteoarticular y endocrinológica, así como los cuestionarios seleccionados para su medición en cada una de las patologías.

En el siguiente capítulo, Capítulo 2, se presentan las hipótesis y objetivos de esta tesis. En el Capítulo 3, se describirá la metodología general empleada para llevar a cabo los estudios de validación y/o reducción de cuestionarios de CVRS. Teniendo en cuenta que algunos de los análisis estadísticos para llevar a cabo estos estudios de validación de los diferentes cuestionarios seleccionados son comunes, este Capítulo 3 de metodología lo plantearemos de manera común. En los Capítulos 4, 5, 6 y 7, se presentan los resultados obtenidos de la aplicación de las diferentes técnicas estadísticas planteadas a los cuestionarios seleccionados. Todos estos resultados están publicados como artículos originales en revistas científicas internacionales, concretamente en cuatro publicaciones, con lo que se presentan los resultados ya publicados. En el Capítulo 8, presentamos una discusión general, recomendaciones de actuación, limitaciones y líneas de investigación futuras. Y finalmente, en el Capítulo 9, presentaremos las conclusiones finales.

Como ya hemos mencionado, los Capítulos 4, 5, 6 y 7 de la presente tesis los dedicamos a la presentación de los resultados, a través de cuatro publicaciones en

revistas científicas internacionales. De estas forma, estos cuatro capítulos dedicados a los resultados están escritos en lengua inglesa, tal y como fueron publicados los artículos originales. El resto de la tesis, está escrita en lengua castellana. En cada caso, se ha respetado la gramática de cada lengua. Por otro lado, hay que puntualizar que la bibliografía correspondiente a cada una de las cuatro publicaciones no ha sido insertada en su correspondiente capítulo de resultados, ya que esta se repite en algunas ocasiones, con lo que se ha unificado junto con el resto de bibliografía de los demás capítulos, y se presenta al final de la tesis de forma conjunta.

1.2. Los cuestionarios de calidad de vida relacionada con la salud

La medida de la salud y de la CVRS ha adquirido una enorme importancia como forma de medir y evaluar resultados de los programas e intervenciones sanitarias. Para ello es necesario disponer de instrumentos de medición de la salud y de la CVRS para que los investigadores, clínicos y decisores sanitarios los puedan utilizar en su práctica diaria (Badia y Alonso, 2007). El método clásico para determinar y evaluar de una forma razonable el impacto de la enfermedad en la vida diaria del individuo y en la sensación de bienestar es la administración de cuestionarios. Algunos aspectos psicológicos de la CVRS tienen definiciones claras, precisas y universalmente acordadas, sin embargo, muchos otros aspectos no son directamente medibles. Estos conceptos constituyen modelos psicológicos que se describen habitualmente como *constructos*, *rasgos latentes*, *factores*, o *dimensiones*. Estas variables latentes o constructos no son directamente medibles, por lo que normalmente se evalúan a través de cuestionarios compuestos de múltiples ítems o preguntas (Fayers y Machin, 2007).

Un cuestionario diseñado para medir CVRS debe basarse en el paciente como fuente de información, reflejando su opinión. Hay que tener en cuenta algunas consideraciones cuando se pretende administrar el cuestionario en la práctica clínica. El cuestionario a administrar al paciente debe ser un cuestionario aceptado por los pacientes, los profesionales sanitarios y los investigadores. Por tanto, al

seleccionar el/los cuestionarios de CVRS a utilizar en un estudio es fundamental maximizar la información recogida y minimizar la carga para investigadores y pacientes. Como norma general, se deben seleccionar cuestionarios que hayan sido evaluados y/o validados debidamente.

Los cuestionarios se pueden clasificar de diferentes maneras. Por un lado, los cuestionarios de CVRS pueden ser autoadministrados o realizarse mediante entrevista personal o telefónica (heteroadministrados), aunque en general un cuestionario que pueda ser completado por el paciente suele resultar más práctico (Sanz, 1991). Los cuestionarios de CVRS se pueden también clasificar en función de si se desea explorar las distintas dimensiones o aspectos del daño que ocasiona una enfermedad, o de si se pretende integrar las dimensiones en un único indicador que resuma y cuantifique las consecuencias de padecer una determinada enfermedad. El primer caso es un enfoque multidimensional, mientras que el segundo sería unidimensional. De esta manera, los cuestionarios unidimensionales están diseñados de tal manera que todos los elementos o ítems se combinan entre sí, por ejemplo, mediante un promedio o una suma, para producir una puntuación global del constructo a medir. Sin embargo, la mayoría de los instrumentos de CVRS son multidimensionales, de forma que están diseñados para agrupar los ítems en escalas separadas correspondientes a los diferentes constructos. De todas formas, aunque muchos investigadores están de acuerdo en que la CVRS es un constructo multidimensional, como hay muchas dimensiones potenciales, no suele ser práctico intentar medir todas ellas simultáneamente en un mismo instrumento (Fayers y Machin, 2007). Aunque la evaluación de la CVRS debe incluir todas las áreas de la vida impactadas por la enfermedad o su tratamiento, la física, la psicológica y la social, el problema reside en la falta de un instrumento único y global capaz de acomodar todos los componentes que definen el concepto de CVRS. Por este motivo, en muchas ocasiones nos encontramos con la necesidad de utilizar varios instrumentos de CVRS en un mismo estudio.

Otra clasificación de los instrumentos de CVRS es en cuestionarios genéricos y específicos (Patrick y Deyo, 1989). Se denominan instrumentos genéricos aquellos

que miden múltiples dimensiones y están diseñados para su aplicación en una gran variedad de individuos. Estos instrumentos están diseñados para uso general, independientemente de la enfermedad o condición del individuo, incluso, a menudo pueden ser aplicables a las personas sanas. Son instrumentos destinados a cubrir una amplia gama de condiciones y tienen la ventaja de que permiten comparar las puntuaciones de individuos de diversas enfermedades entre si y con la población general. Por otra parte, estos instrumentos no logran enfocarse en los temas de especial interés para los pacientes con la enfermedad, y muchas veces no tienen la sensibilidad para detectar diferencias que se producen como consecuencia de un tratamiento. Esto ha llevado al desarrollo de cuestionarios específicos de la enfermedad (Fayers y Machin, 2007). De esta forma, los instrumentos específicos se centran en la medición de aspectos concretos de una determinada enfermedad, de una población, o un aspecto clínico. En general, tienen mayor poder de discriminación (más capaces de detectar diferencias entre tratamientos). Sin embargo, si se pretende comprobar el impacto que una determinada intervención tiene en el estado de salud, hay que tener en cuenta que esta también influye en las características más generales del paciente. Por tanto, al evaluar el estado de salud será conveniente utilizar también algún cuestionario genérico (Guyatt y cols., 1993). Por otro lado, si se pretende dar respuesta a problemas específicos de salud, se buscan instrumentos que sean sensibles a variaciones de CVRS en enfermedades específicas. Como normal general, es aconsejable incluir en el estudio tanto instrumentos específicos como genéricos.

1.3. Atributos y propiedades psicométricas de los cuestionarios

La MOT enfatiza la necesidad de expandir la disponibilidad y el uso de cuestionarios autoadministrados o en formato entrevista (heteroadministrados), diseñados para evaluar la salud y los resultados del cuidado de la salud desde el punto de vista del paciente. Para cumplir este fin, la MOT consideró conveniente identificar los cuestionarios creados hasta el momento, indexarlos en una librería de cuestionarios, y así poderlos diseminar entre las personas o entidades interesadas en su uso. Esta Fundación decidió crear un Comité de Consejo

Científico (Scientific Advisory Comittee, SAC) que se encargaría de revisar y evaluar la aceptabilidad de los cuestionarios que luego fuesen a ser distribuidos por la MOT. Para realizar esta revisión de cuestionarios, la SAC determinó los atributos y criterios que iba a emplear en la revisión de los cuestionarios (Aaronson y cols., 2002). Los atributos y criterios que consideraron fueron los siguientes:

1) el modelo conceptual y de medida;
2) la interpretabilidad;
3) la carga;
4) la forma de administración;
5) la adaptación cultural y traducción; y
6) las propiedades psicométricas.

Por tanto, estos criterios son los que debemos tener en cuenta de manera general a la hora de seleccionar un cuestionario. Otros grupos tales como el "European Regulatory Issues on Quality of Life Assessment", la "International Society for Pharmaeconomics and Outcomes Research", y las agencias reguladoras tales como la "United States Food and Drug Administration" y la "European Medicines Agency", también apoyan estos criterios a tener en cuenta en la selección de instrumentos de medida de la CVRS (Huang y Speight, 2013).

El *modelo conceptual y de medida* se refiere a la justificación y descripción del concepto o conceptos a medir por el cuestionario, y de la población o poblaciones a la cual está destinada dicha medición, así como a la relación entre ambas. El modelo de medida consiste en la puesta en marcha del modelo conceptual, representando este modelo conceptual en la escala del instrumento y en su estructura, y en el procedimiento seguido para crear las puntuaciones de la escala y/o de las subescalas. La idoneidad de un modelo de medida se contrasta examinando la evidencia de si (1) la escala mide un único dominio conceptual o constructo, (2) múltiples escalas miden distintos dominios, (3) la escala representa adecuadamente la variabilidad en el dominio, y (4) se justifica el nivel de medida elegido en la escala y su procedimiento de puntuación.

La *interpretabilidad* se refiere al grado en que podemos interpretar fácilmente las puntuaciones de las escalas del instrumento. Un cuestionario ha de proporcionar información sobre como han de ser presentados los resultados de dicho cuestionario. Se trata de en que grado podemos con facilidad darle sentido a la puntuación que se obtiene en el instrumento (Testa, 2000). La existencia de parámetros o valores de referencia suele ayudar a la interpretación de los resultados que obtenemos de un cuestionario. Por ejemplo, el hecho de proporcionar datos comparativos de la distribución de las puntuaciones en una variedad de grupos poblacionales, incluso en población general cuando sea posible, resulta de gran utilidad a la hora de interpretar los resultados de las puntuaciones. Los resultados obtenidos por el mismo instrumento en otros estudios también pueden resultar de gran utilidad. Otra información que puede ser de utilidad para la interpretabilidad del cuestionario es la relación entre las puntuaciones del cuestionario y ciertas condiciones clínicas o determinados tratamientos o intervenciones de reconocida eficacia.

La *carga* del instrumento se refiere a la inversión de tiempo, de esfuerzo y otras demandas causadas por la compleción del cuestionario que la persona que responde ha de hacer, o bien el entrevistador en caso de cuestionarios heteroadministrados. Este aspecto también es importante tenerlo en cuenta ya que cuando un cuestionario es complejo con ítems difíciles de entender, o con un número elevado de preguntas, la tasa de respuesta tiende a ser más baja.

La *forma de administración* del instrumento se refiere a si el cuestionario es autoadministrado o heteroadministrado. En algunas ocasiones puede resultar que un instrumento puede ser administrado de diferentes maneras, en cuyo caso, debiera de haber evidencia de la interpretabilidad, carga y propiedades psicométricas de cada modo de administración, así como una comparativa de los diferentes modos de aplicación del cuestionario.

La *adaptación cultural y traducción* se refiere al grado en que un instrumento está adaptado y traducido para su aplicación en otras culturas y lenguas. El proceso de

traducción y adaptación de un cuestionario se debe llevar a cabo de manera tan rigurosa como su propio desarrollo, para así evitar la introducción de cualquier error en el cuestionario y en los matices que pueden afectar la forma en que los pacientes respondan a los ítems. El objetivo del proceso de adaptación y traducción debe ser garantizar que todas las versiones del cuestionario sean igualmente claras, precisas y equivalentes en todos los aspectos al cuestionario original (Fayers y Machin, 2007).

Y por último, dentro de los atributos que la SAC considera que ha de cumplir un instrumento, y probablemente el más importante, son las *propiedades psicométricas*, refiriéndose a la fiabilidad, validez y sensibilidad al cambio de un instrumento. Toda medida, desde la medición de la presión sanguínea hasta la medición de la CVRS, debe satisfacer propiedades básicas para que puedan ser de utilidad en la práctica clínica. Estas propiedades son ante todo la validez, fiabilidad, y sensibilidad a cambios significativos. La *psicometría* es la ciencia para evaluar estas características de medida de las escalas (Staquet y cols., 1998), que nos permite medir constructos intangibles (como puede ser la CVRS de un individuo) a través de los indicadores tangibles (ítems) más adecuados (Streiner y Norman, 1995). La validez, fiabilidad y sensibilidad al cambio aunque están interrelacionados, cada una es importante de manera independiente. De manera general, la fiabilidad se refiere a la precisión de la escala de medida (mínimo error de medida); la validez se refiere a la capacidad del instrumento de medir aquellas características que pretende medir y no otras; y la sensibilidad al cambio se refiere a la capacidad del instrumento de detectar cambios en la respuesta de un individuo a lo largo del tiempo (responsiveness) (Donovan y cols., 1989). A su vez, cada una de estas propiedades psicométricas se puede subdividir en diferentes subtipos, tal y como se muestra en la Figura 1.1, y que introduciremos con mayor detalle a continuación en los siguientes apartados. La evaluación de estas características es una tarea compleja y sin fin, en especial la evaluación de la validez. En la investigación de la CVRS, el proceso de validación consiste en acumular más y más evidencia de que las escalas son sensibles y que se comportan de la manera que se

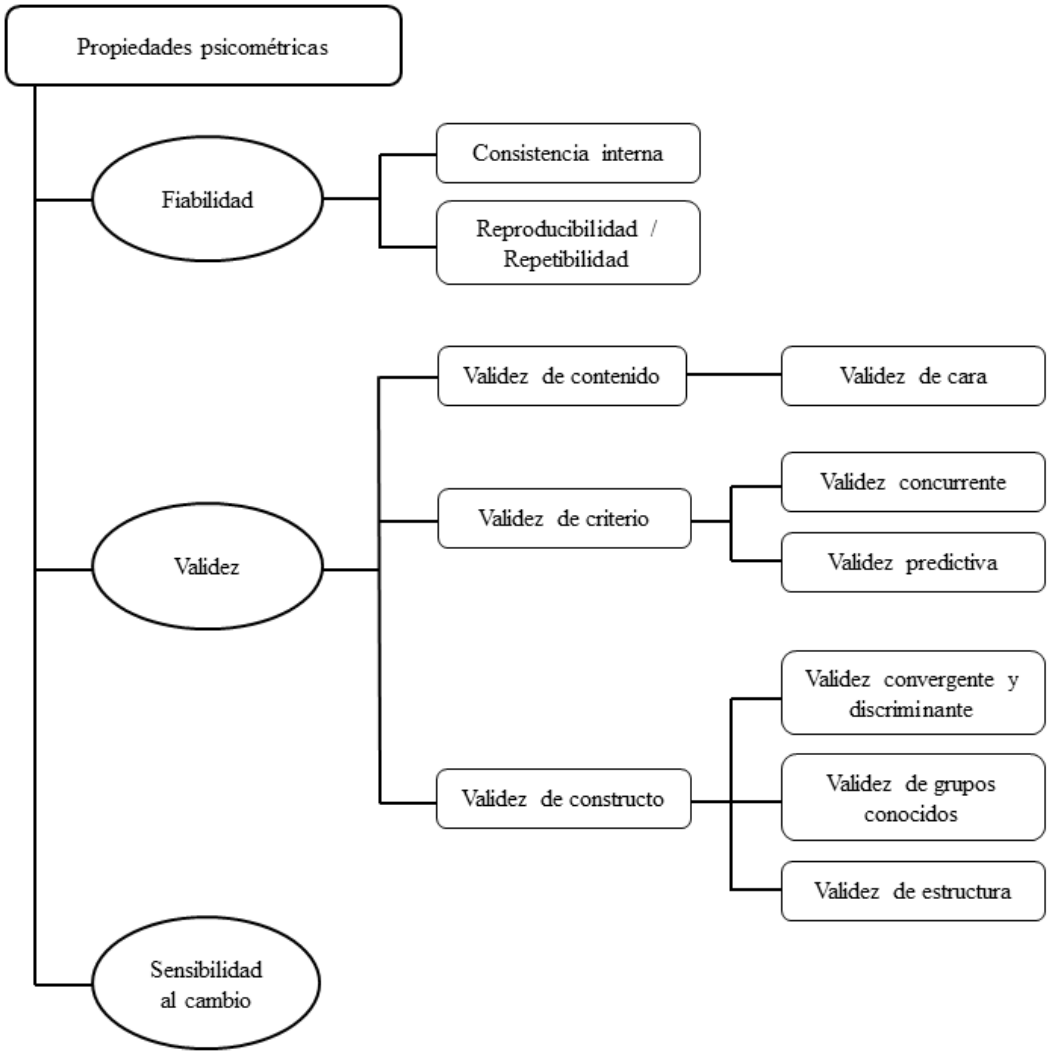


Figura 1.1. Clasificación de las propiedades psicométricas.

anticipa. Por tanto, la validación de un instrumento es un proceso para determinar si existen motivos para pensar que el instrumento mide lo que se pretende medir, y que es útil para los fines previstos. Este proceso de validación consiste en un número de etapas, en las que se espera recoger evidencia convincente de que el instrumento cubre los constructos para los que fue diseñado y que produce mediciones útiles que reflejan la CVRS de los pacientes (Fayers y Machin, 2007).

1.4. La fiabilidad

La fiabilidad de un instrumento de medición de la CVRS indica la estabilidad de los resultados cuando se repite el proceso de medición en circunstancias similares. Es decir, la fiabilidad de una escala de medida se basa en examinar si se reproducen los mismos resultados de manera consistente cuando no se han producido cambios en la situación del individuo. Dicho de otra manera, la fiabilidad de una escala viene establecida por el grado en que la medición está libre de error aleatorio. Idealmente, los pacientes cuya calidad de vida no ha cambiado, deberían dar respuestas muy similares y repetibles cada vez que estas son evaluadas. Si hay una considerable variabilidad aleatoria, las mediciones no son fiables. Resultaría difícil saber cómo interpretar los resultados individuales de un paciente si las mediciones no son fiables. Una baja fiabilidad puede a veces ser un aviso del incumplimiento de la validez, y de que la medición está detectando algo diferente de lo que pretende medir (Fayers y Machin, 2007). La evaluación de la fiabilidad se realiza mediante la medición de (a) la consistencia interna y (b) la reproducibilidad.

La *consistencia interna* hace referencia a la homogeneidad de los ítems que componen una escala. Si los distintos ítems de un cuestionario pretenden medir un mismo concepto es esperable que las respuestas a estos ítems estén relacionadas entre sí (Badia y Alonso, 2007). La *reproducibilidad* o *repetibilidad* de las mediciones se refiere al grado en que la medida proporciona resultados reproducibles o similares si se usa repetidamente en los mismos individuos y cuando la condición del individuo no ha cambiado. La repetibilidad puede referirse a medidas repetidas en el tiempo, dando así lugar a la fiabilidad test-retest, o bien a medidas repetidas por diferentes observadores dando así lugar a la fiabilidad inter-observador. Si un paciente se encuentra en una condición estable, el instrumento debería proporcionar resultados repetibles y reproducibles si se usa repetidamente en el paciente. Esto se suele evaluar utilizando un estudio test-retest, con pacientes que están estables en su enfermedad y que no se espera tengan cambios. Cuando el instrumento tiene que ser aplicado por entrevista, la fiabilidad inter-observador toma importancia. Es un indicador de la estabilidad de

las mediciones cuando el proceso de medición es realizado por varios entrevistadores sobre el mismo individuo. Trata de ver si independientemente del entrevistador el instrumento proporciona resultados similares (Badia y Alonso, 2007; Fayers y Machin, 2007).

1.5. La validez

De las características de un instrumento de CVRS, probablemente la validez es la característica más importante que se debe exigir a un buen cuestionario. La validez, se define como el grado en que un instrumento mide lo que se propone medir, y que es útil para ese propósito (Badia y Alonso, 2007; Streiner y Norman, 1995). En el ámbito de la CVRS los conceptos que se pretenden medir no son objetos físicos que permitan una medición directa. Se trata de conceptos o constructos que evaluamos indirectamente mediante un sistema de indicadores. No existe una forma única de evaluar la validez de un instrumento de medición y suele distinguirse en tres aspectos principales: validez de contenido, validez de criterio, y validez de constructo. La validez de contenido se refiere a la medida en que los elementos son sensibles y reflejan el dominio previsto de interés. La validez de criterio considera si la escala tiene asociación empírica con criterios externos, como otros instrumentos establecidos. La validez de constructo examina la relación teórica de los ítems entre sí y con las escalas hipotéticas (American Educational Research Association y cols., 1999; Badia y Alonso, 2007; Eignor, 2001; Fayers y Machin, 2007).

1.5.1. La validez de contenido

La *validez de contenido* trata de proporcionar evidencia de que el contenido del dominio es apropiado para medir el constructo que se pretende medir, y que es una reflexión adecuada del constructo que va a ser medido. Así, se refiere a la adecuación del contenido de un instrumento en términos de la cantidad y el alcance de las preguntas individuales que contiene. Consiste en revisar el instrumento para asegurarse de que parece ser sensible y cubre todas las

cuestiones pertinentes. Por lo tanto, la validación de contenido implica el examen crítico de la estructura básica del instrumento, basándose en el análisis lógico del concepto que se pretende medir, y particularmente, de la definición de las dimensiones que abarca el cuestionario y sus límites con otros conceptos relacionados. Para comprobar la validez de contenido, el diseño y el desarrollo de un instrumento deben seguir una metodología rigurosa (Fayers y Machin, 2007). Sobre la base de este análisis puede determinarse a priori si el instrumento de medición contiene las dimensiones e ítems representativos de todos los elementos que forman la definición del concepto y si el número es proporcional a la importancia que concede la teoría a cada una de las dimensiones de la definición (Badia y Alonso, 2007).

La cobertura comprensiva es uno de los aspectos más importantes de la validez de contenido, y todos los aspectos relevantes deberían estar cubiertos por el instrumento. Por ejemplo, un instrumento diseñado para medir sintomatología debería incluir ítems referentes a todos los síntomas más relevantes de la enfermedad. Si no, podríamos encontrarnos con la no detección de diferencias entre grupos de pacientes. El grado de la cobertura de los ítems no es susceptible de pruebas estadísticas formales, y depende principalmente de asegurarnos que el instrumento ha sido desarrollado siguiendo una metodología rigurosa. El proceso de generación de los ítems del cuestionario debe incluir opiniones de expertos de la enfermedad, revisión de la literatura, y entrevistas con pacientes que padecen dicha enfermedad. Al mismo tiempo, todos los ítems incluidos en el instrumento deben ser relevantes para el concepto que se pretende medir, y todo ítem irrelevante debería ser excluido. Los ítems también deberían ser excluidos en caso de ser redundantes, ya que únicamente duplicaríamos la información ya contenida en otros ítems (American Educational Research Association y cols., 1999; Eignor, 2001).

La *validez de cara* (face validity) consiste en comprobar si los ítems en un instrumento o una dimensión concreta cubren el área que se pretende cubrir de forma clara y precisa. Requiere la comprobación de si los ítems en un instrumento

"parece" aparentemente que cubren los temas destinados a cubrir, claramente y sin ambigüedades. La validez de cara está estrechamente relacionada con la validez de contenido y, a menudo se considera que es un aspecto de ella. La principal distinción es que la validez de cara se refiere a la revisión crítica de un instrumento después de que se ha construido, mientras que la mayor parte de la validación del contenido consiste en asegurar que los procedimientos de desarrollo fueron seguidos rigurosamente y documentados (Fayers y Machin, 2007).

1.5.2. La validez de criterio

La *validez de criterio* se refiere a que el cuestionario sea adecuado al problema que se quiere medir. Esto se comprueba proporcionando evidencia sobre el nivel en que las puntuaciones del instrumento están relacionadas con una medida criterio o variable externa. Las medidas criterio son aquellos cuestionarios o herramientas de medida ya creadas que miden el constructo de nuestro interés y que han sido ampliamente aceptadas como herramientas validas de medida (American Educational Research Association y cols., 1999; Eignor, 2001). Normalmente, cuando existe un elevado consenso y tradición entre los investigadores sobre un buen procedimiento de medida, este se considera un estándar o patrón de oro (gold estandar) con el que se compararán todos los nuevos instrumentos (Badia y Alonso, 2007).

En el área de la auto-evaluación del estado de salud y de la medición de la CVRS, la validez de criterio no se suele proporcionar ya que no hay un instrumento de medida de la salud que sea ampliamente aceptado en el panorama científico. Como no existe "estándar oro" en los instrumentos de CVRS, la mejor aproximación que se puede hacer es comparar el instrumento con algún otro validado y ampliamente aceptado. Esto puede ser razonable en especial cuando el objetivo es analizar la validez de criterio de una reducción de un instrumento. En este caso, la medida criterio sería la versión larga de dicho instrumento (American Educational Research Association y cols., 1999; Eignor, 2001). Más frecuentemente, la

justificación de crear una nueva herramienta es que los investigadores creen que la existente no es óptima y adecuada. En este caso, la comparación de la nueva con la ya existente tiene también un valor limitado, ya que la existente, en efecto, ha sido rechazada como estándar oro. Otra aproximación puede ser el utilizar métodos indirectos de comparación. Una entrevista detallada, con personal capacitado en técnicas de entrevista, podría producir estimaciones de los constructos que se pretenden medir (Fayers y Machin, 2007).

La validez de criterio se puede subdividir en dos tipos: validez concurrente y validez predictiva. Cuando la relación del instrumento con la medida criterio se establece en el mismo momento temporal se dice que se está estudiando la *validez concurrente*. En ocasiones, la variable de criterio es un acontecimiento futuro que se intenta predecir mediante el resultado del instrumento de medición. En esta circunstancia la validez analizada se denomina *validez predictiva* (Badia y Alonso, 2007). La validez predictiva se refiere a la habilidad del instrumento para predecir un estado de salud futuro, un evento futuro o un resultado del test futuro. Por ejemplo, tal y como dicen Fayers y Machin (2007), frecuentemente se ha dicho que las puntuaciones globales de calidad de vida son predictivas del tiempo de supervivencia en pacientes con cáncer, y que la evaluación de la CVRS proporciona información pronóstico adicional complementaria a medidas objetivas, tales como el grado del tumor o la extensión de la enfermedad. La implicación de esto es que ese estado de salud futuro puede servir como criterio contra el cual podemos comparar el instrumento.

1.5.3. La validez de constructo

Entre las diferentes maneras de proporcionar evidencia de la validez de un instrumento, una de las más importantes es la *validez de constructo*, que consiste en proporcionar evidencia de que la forma de interpretar las puntuaciones del instrumento es correcta según la teoría y los constructos que se miden (Fayers y Machin, 2007). El tema de la validez de constructo es difícil y controvertido. Se trata de formar primero un modelo hipotético, que describe los constructos que se

están evaluando y postular sus relaciones. Una vez recogidos los datos, se realiza una evaluación para ver en qué medida se confirman estas relaciones. Si los resultados confirman las expectativas previas sobre los constructos, la implicación es que el instrumento puede ser válido y que por lo tanto puede ser utilizado para hacer inferencias acerca de los pacientes.

Más formalmente, la validez de constructo abarca una variedad de técnicas, todas ellas enfocadas a evaluar dos aspectos: primero, si el constructo teórico postulado es un modelo adecuado; y segundo, si las escalas de medida se corresponden al constructo postulado. Traduciendo esto a la práctica, la validez de constructo se basa primordialmente en comprobar (Fayers y Machin, 2007):

a) Dimensionalidad: ¿todos los ítems de una escala se relacionan con una única variable latente, o existe evidencia de que se necesitan más variables latentes para explicar la variabilidad observada?
b) Homogeneidad: ¿todos los ítems de una subescala parecen tener el mismo peso sobre la variable latente?
c) Solapamiento entre variables latentes: ¿algunos ítems de una subescala correlacionan con otras variables latentes?

La validez de constructo es un proceso largo y continuo, y además, de los diferentes tipos de validez que hemos mencionado, la validez de constructo es el más susceptible de exploración mediante el análisis numérico y estadístico (Fayers y Machin, 2007). Dentro de la validez de constructo, podemos incluir aspectos tales como la validez convergente y discriminante, la validez de grupos conocidos o la validez de estructura, que se describen a continuación.

Validez convergente y discriminante

En el ámbito de la CVRS encontramos numerosos conceptos que establecen entre ellos interrelaciones complejas. Por ello, no ha de extrañar que los instrumentos de medición que pretenden evaluar diferentes dimensiones de la CVRS mantengan entre ellos un cierto grado de relación. Partiendo de este hecho, la *validez*

convergente y discriminante pretende analizar si una matriz de relaciones entre diferentes instrumentos es coherente con lo esperable a partir del conocimiento teórico. En concreto, cabe esperar que dos instrumentos que midan el mismo concepto establezcan interrelaciones elevadas entre sus dimensiones o componentes, e interrelaciones más débiles con las puntuaciones de instrumentos que evalúan otras dimensiones menos relacionadas (Badia y Alonso, 2007). De esta forma, la *validez convergente* (algunas veces también denominada validez concurrente) consiste en mostrar que una dimensión de CVRS correlaciona apreciablemente con todas las demás dimensiones con las que en teoría miden el mismo concepto o similar. De forma similar, la *validez discriminante o divergente*, reconoce que algunas dimensiones de CVRS están prácticamente incorrelacionadas con otras, y por lo tanto, las correlaciones entre estas escalas serán bajas (Fayers y Machin, 2007).

Validez de grupos conocidos

Una de las formas más simples de la validez de constructo es la *validez de grupos conocidos*. En muchas ocasiones un instrumento de medida de salud se utiliza para discriminar entre grupos de individuos o de pacientes con distintos niveles de gravedad de una afección determinada. Resulta fundamental demostrar esta capacidad de diferenciación en instrumentos de medición de la calidad de vida que pretendan tener capacidad discriminante. Por tanto, se trata de que el instrumento reproduzca resultados diferentes al ser aplicado a diferentes grupos de pacientes (Badia y Alonso, 2007). Se basa en el principio de que cierto grupo específico de pacientes tendrá una puntuación diferente a otros, y que el instrumento debería ser sensible a estas diferencias. Por ejemplo, se espera que los pacientes con cáncer avanzado tengan una peor CVRS que los pacientes con enfermedad temprana. Una escala valida debería mostrar diferencias entre estos dos grupos de pacientes. Una escala que no puede distinguir de manera satisfactoria entre grupos de pacientes con conocidas diferencias, bien porque tiene falta de sensibilidad o bien porque proporciona resultados contrarios a lo esperado, es poco probable que sea de valor para muchos fines (Fayers y Machin, 2007).

Validez de estructura

Dentro de la validez de constructo podemos decir que la *validez de estructura* es la que juega un papel más importante, siendo también el tema con más controversia. La validez de estructura trata de la evaluación del grado en el que un instrumento mide el constructo para el cual fue diseñado. Para establecer la validez de estructura de un instrumento se realiza un análisis interno del mismo para conocer la estructura básica subyacente (los constructos no observables, o factores latentes) de forma que las relaciones existentes entre los ítems del cuestionario se expliquen por la existencia de dichos factores (Fayers y Machin, 2007). Se trata de comprobar que la estructura hipotética subyacente del instrumento es correcta y adecuada.

Como ya hemos comentado, la validez de constructo, y en particular la validez de estructura, es la más susceptible de exploración por el análisis numérico y estadístico. Además, existen diferentes perspectivas y teorías de enfoque, desde la teoría clásica del test (TCT) hasta la teoría de la respuesta al ítem (TRI), e incluso dentro de cada una de las teorías, se proponen diferentes técnicas. Sin embargo no hay un consenso sobre que metodología utilizar, y es en este aspecto donde sobre todo centraremos este trabajo de tesis doctoral. Por ello, la siguiente subsección 1.5.4 se destina íntegramente a la descripción de las diferentes teorías de medida y los diferentes métodos para la evaluación de la validez de estructura.

1.5.4. Métodos para la evaluación de la validez

Una teoría de la medición es una teoría sobre cómo las puntuaciones generadas de los ítems representan el constructo que se pretende medir (de Vet y cols., 2011). Esta definición sugiere que las teorías de medición se aplican a instrumentos compuestos por muchos ítems. Sin embargo, hay que tener en cuenta que las teorías de medición no siempre son necesarias para instrumentos multi-ítem, sólo son necesarias cuando miden un constructo no observable. Para características observables, suele ser obvio qué ítems contribuyen al constructo que se pretende

medir y normalmente no hacen falta teorías de medición. Por ejemplo, las comorbilidades, que se caracteriza por el número de enfermedades, el tipo de enfermedades y su severidad, es algo observable. Sin embargo, si hablamos de la carga que supone las comorbilidades, esto ya se va a algo menos observable. Por tanto, medir constructos no observables es todo un reto. Estos constructos son muy frecuentes en disciplinas como la psicología o la psiquiatría, y también en instrumentos de CVRS o de resultados percibidos por los pacientes en disciplinas médicas. Estos constructos son normalmente medidos a través de múltiples ítems directamente observables. Entonces, necesitamos teorías de medición que describan las relaciones estadísticas entre los ítems y los constructos. Así, introducimos la manera gráfica de representar estas relaciones, tal y como se muestra en la Figura 1.2 en las que mediante círculos se representan los constructos no observables directamente y mediante rectángulos los ítems directamente observables.

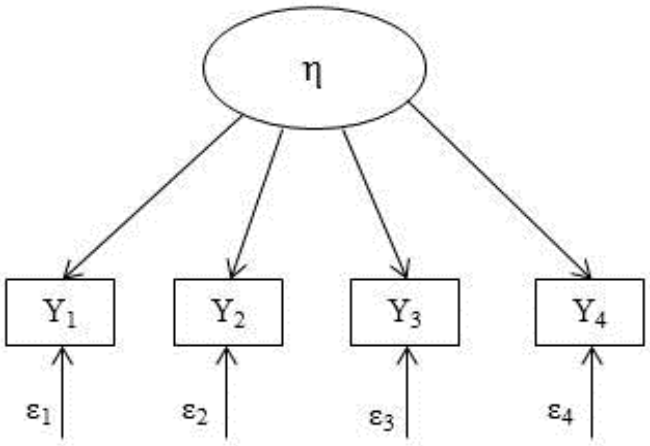


Figura 1.2. Representación gráfica de un modelo de medición (de Vet y cols., 2011).

Existen dos teorías de medida: la TCT y la TRI (de Vet y cols., 2011). La TCT se desarrolló a comienzos del siglo XX por psicólogos tales como Spearman y Cronbach (Lord y Novick, 1968). Se obtiene información del constructo a través de la medición de ítems que son manifestaciones del constructo, ya que los ítems son más fáciles de medir. Estos forman lo que se denomina "psicometría clásica" y se

basa en gran parte en escalas que se obtienen sumando los ítems que la componen, o en modelos lineales de dichos ítems, tales como modelos de análisis factorial (Fayers y Machin, 2007).

Dentro de la psicometría clásica, tradicionalmente se ha empleado el Análisis Factorial Exploratorio (AFE), que trata de identificar el número de factores latentes y la naturaleza de cada uno de ellos, explorando patrones entre las correlaciones de los ítems (Fayers y Machin, 2000; Kim y Mueller, 1978a y 1978b). Otro método también clásico aunque quizás algo menos utilizado, es el Análisis Factorial Confirmatorio (AFC), que trata de contrastar hipótesis realizadas *a priori* sobre la estructura y relación entre las variables medidas (ítems) y los factores latentes (Fayers y Machin, 2000; Hatcher, 1994; Long, 1983). Mientras que el AFE es un método meramente exploratorio, el AFC es un método confirmatorio para testar una estructura hipotetizada a través de índices de bondad de ajuste.

Otros métodos dentro de la psicometría clásica, aunque más sofisticados y bastante menos utilizados en el ámbito de cuestionarios de CVRS, son los Modelos de Ecuaciones Estructurales (MEE). Los MEE son una generalización de los modelos factoriales que permiten ajustar estructuras mucho más complejas, además de permitir ajustar modelos causales que los métodos anteriores no son capaces de ajustar (Bollen, 1989; Fayers y Machin, 2000; Hatcher, 1994). Por lo tanto, los MEE podrían proporcionar una muy buena alternativa a los métodos más clásicos como el AFE y AFC, cuando estos no son capaces de ajustar la estructura subyacente del cuestionario.

De todas formas, todavía hay algunos aspectos dentro del Análisis Factorial (AF) y los MEE que están sin resolver, son algo más complicados de tratar, y están en fase de desarrollo. Uno de ellos es el supuesto de normalidad. La mayoría de los índices de ajuste de estos métodos se basan en el supuesto de normalidad de la distribución de las variables. Sin embargo, la mayoría de los cuestionarios de CVRS constan de ítems con escalas de respuesta tipo Likert, es decir, variables categóricas y generalmente con distribución no normal (Fayers y Machin, 2007).

La preocupación por el buen uso del AF con datos categóricos no ha pasado desapercibida (Bartholomew, 2007; Bartholomew y cols., 2002). Para solventar este problema se ha desarrollado metodología para la utilización del AF y MEE para datos categóricos (Lee y cols., 1995), aunque su uso no está muy difundido, debido probablemente a que no está implementada en la mayoría de los paquetes estadísticos.

Desde comienzos del siglo XX, la construcción y uso de los test psicométricos se han basado principalmente en la TCT, siendo un modelo simple, flexible y muy conocido (Gulliksen, 1950), pero sin embargo, no está exento de limitaciones (Embretson y Hershberger, 1999; Prieto y Delgado, 2003). Así, en 1960 el matemático danés Georg Rasch (Rasch, 1960) propuso un modelo de medida que permite solventar muchas de las deficiencias de la TCT y construir escalas más adecuadas y eficientes como son los modelos de la TRI, también denominada "psicometría moderna". Éstos son modelos que hacen hincapié en la importancia de los modelos de respuesta al ítem, en el que los individuos con un determinado nivel de capacidad tienen una "probabilidad de responder positivamente" a una pregunta. La TRI fue pionera en educación, donde cada pregunta únicamente tenía dos opciones de respuesta, si/no, o correcto/incorrecto. Utilizando ítems que abarcaban un amplio rango de dificultad, la habilidad del individuo se puede puntuar con un alto grado de precisión. No podemos distinguir a los alumnos más capaces de los ligeramente menos capaces si todos los ítems son demasiado fáciles. Así, un examen ha de tener alguna pregunta muy difícil que nos permita discriminar a este nivel. De la misma manera, son necesarias algunas preguntas fáciles para distinguir entre los estudiantes menos hábiles (Fayers y Machin, 2007).

La TRI se basa en que un individuo con un nivel concreto de calidad de vida o de funcionamiento tendrá una cierta probabilidad de responder de forma positiva a una pregunta. Esta probabilidad dependerá de la dificultad de la pregunta. Por ejemplo, muchos de los pacientes con cáncer responderán "Si" a preguntas "fáciles" como "¿Tiene algo de dolor?", pero sólo los pacientes con un alto nivel de dolor

responderán "Si" a preguntas más "difíciles" como "¿Has tenido mucho dolor?". Por tanto, la hipótesis subyacente en la TRI es que la respuesta a un ítem de un cuestionario depende del nivel de dificultad que el ítem presente y el nivel de habilidad respecto al constructo del individuo que responde (Linacre, 2005; Tang y cols., 2007; Tennant y cols., 2004). Así, los modelos de la TRI son típicamente utilizados para medir la habilidad del paciente, por ejemplo, la habilidad física o cognitiva. Para explicar los modelos de la TRI, pongamos el ejemplo de la "capacidad de andar" de un grupo de pacientes. Asumimos que es un constructo unidimensional, que podrá tener un rango desde incapaz de andar hasta andar sin ninguna limitación. Cada paciente tendrá una posición o localización en este continuo de "capacidad para andar", la cual se denomina *localización del paciente* o *habilidad del paciente*. Los modelos de la TRI hacen posible estimar estas localizaciones de los pacientes a través de sus puntuaciones en el conjunto de ítems. Algo típico de los modelos de la TRI es que los ítems también tienen su localización en esa misma escala de capacidad o habilidad para andar. Esta localización se denomina *localización del ítem* o *dificultad del ítem*. Las medidas basadas en los modelos de la TRI, nos permiten obtener información tanto sobre la localización de los ítems como de los pacientes. Así, los modelos de la TRI describen la asociación entre el nivel de habilidad o severidad del respondedor sobre el constructo a medir y la probabilidad de una respuesta concreta al ítem, de forma que la dificultad del ítem y la habilidad del individuo son aspectos que están vinculados. Cuanta más habilidad tenga el paciente mayor probabilidad de que de una respuesta positiva a cualquier pregunta relevante. Cuanto mayor sea la dificultad del ítem, menos probable que ese ítem sea respondido de manera positiva por la mayoría de los pacientes (de Vet y cols., 2011). Por tanto, dentro de la TRI, los ítems tienen que variar en cuanto al nivel de dificultad que presentan, y se asume, que los pacientes tendrán diferente probabilidad de responder de manera positiva a una pregunta, de acuerdo con su nivel de habilidad (nivel de habilidad en relación a la variable latente) (Fayers y Machin, 2007).

Los modelos de la TRI son modelos matemáticos que se basan en que la probabilidad de respuesta positiva a un ítem depende de la habilidad del individuo

con respecto al constructo que se mide, y de la dificultad que presente el ítem (Cella y cols., 2002). Se postula un rasgo subyacente no observable en el cual los ítems se ordenan de manera jerárquica desde el más fácil hasta el más difícil. Los modelos de la TRI incluyen un conjunto de asunciones que hay que tener en cuenta. Una asunción es la unidimensionalidad, ya que asume que los ítems de la escala son de una única dimensión o constructo.

Dentro de los modelos de la TRI, los primeros modelos fueron desarrollados para ítems de respuestas dicotómicas. Los tres modelos de la TRI para datos dicotómicos más frecuentes son los modelos logísticos de 1, 2, y 3 parámetros, así llamados dependiendo del número de parámetros que incluyan. El hecho de que sean de 1, 2 o 3 parámetros depende del número de parámetros que se utilicen en la función que modeliza la relación entre el constructo latente y las respuestas a los ítems. Los *modelos logísticos de 1–parámetro* se centran en el nivel de dificultad del ítem como único parámetro de interés. Estos modelos también se denominan modelos Rasch debido a su inventor, Georg Rasch (Rasch, 1960). Los *modelos logísticos de 2–parámetros* incluyen además de la dificultad del ítem, también la discriminación del ítem. La discriminación del ítem se define como la magnitud del ítem en relación al dominio de CVRS subyacente medido por la escala (Huang y Speight, 2013). Y los *modelos de 3–parámetros* incorporan un tercer parámetro de incertidumbre (guessing parameter) o también llamado parámetro de oportunidad (chance-parameter). Este parámetro indicaría que una persona con muy bajo nivel en el rasgo latente respondería positivamente por mero azar (Hambleton y cols., 1991).

También se han desarrollado extensiones de estos modelos de la TRI para manejar ítems de respuesta politómica. Masters y Wright (Cella y cols., 2002) hicieron una revisión de las extensiones del modelo Rasch de 1–parámetro, incluyendo el *Partial Credit Model* (Masters, 1982) y el *Rating Scale Model* (Andrich, 1978) como modelos adecuados para ítems de respuesta tipo Likert. Dentro de los modelos TRI de 2–parámetros generalizado a datos politómicos ordenados tenemos el *Graded Response Model* (Samejima, 1969) y el *Generalized Partial Credit Model* (Muraki,

1992). Ambos modelos son adecuados para ítems tipo Likert, con opciones de respuesta ordenadas, e incluyen tanto el parámetro de dificultad como el de discriminación de los ítems (DeMars, 2010; Hambleton y cols., 1991). Para respuestas nominales está el *Nominal Response Model* desarrollado por Bock (1972) como una generalización del modelo de 2–parámetros a variables politómicas (Hambleton y cols., 1991).

Así, para el estudio de la validez de estructura, se han empleado diferentes técnicas estadísticas, desde la psicometría clásica basada en modelos de análisis factorial, hasta la psicometría más moderna basada en los modelos probabilísticos de la TRI.

1.6. La sensibilidad al cambio

Otra de las características psicométricas que ha de cumplir un cuestionario es la *sensibilidad al cambio* (responsiveness). Se trata de la capacidad del instrumento para detectar el cambio (Deyo y cols., 1991; Husted y cols., 2000). Un instrumento de CVRS no sólo debe ser fiable, proporcionando resultados reproducibles cuando la condición del paciente es estable y no cambia, sino además, debe responder a cambios relevantes en la condición del paciente. Si la progresión de la enfermedad causa una mejoría o un deterioro en la CVRS del paciente, esperamos que la medición del instrumento responda en consecuencia. Por tanto, se trata de averiguar si el cuestionario puede detectar diferencias en los resultados, incluso cuando estas diferencias son pequeñas (Terwee y cols., 2001). El instrumento de medida debería ser suficientemente sensible al cambio para detectar cambios relevantes cuando sabemos que la condición del paciente ha sido alterada.

La sensibilidad al cambio de un instrumento de medida de la CVRS se evalúa mediante un estudio longitudinal de pacientes en los que esperamos que ocurran cambios, habitualmente en situaciones en las que se realiza una intervención o un tratamiento de reconocida eficacia. El instrumento ha de ser capaz de detectar cambios reales, positivos o negativos, en la salud (Badia y Alonso, 2007). La sensibilidad al cambio se puede entender como la proporción de señal

(entendiendo como el cambio real que ha ocurrido) con respecto al ruido (entendiendo como la variabilidad que ha ocurrido en las puntuaciones entre ambas administraciones pero que no tiene que ver con el cambio en el estado del paciente). Un instrumento de CVRS será de uso limitado a menos que sea capaz de detectar cambios en los pacientes que realmente han tenido un cambio. Los cuestionarios específicos de CVRS tienden a ser más sensibles al cambio que los instrumentos genéricos (Fayers y Machin, 2007).

1.7. Reducción de cuestionarios

Aunque existen cuestionarios de CVRS con buenas propiedades psicométricas, a menudo su longitud limita en gran medida su aplicación, ya que nos encontramos con cuestionarios que requieren de mucho tiempo para ser cumplimentados por el individuo o el entrevistador. Los instrumentos de CVRS más reducidos tienen un mejor cumplimiento de los mismos y una mayor tasa y calidad de respuesta, por lo que disponer de una versión reducida de un cuestionario de CVRS que mantenga las mismas buenas propiedades psicométricas que su versión larga, es de gran utilidad tanto en la práctica clínica como en investigación (Coste y cols., 1997; Moran y cols., 2001; Prieto y cols., 2003).

Idealmente, un cuestionario ha de ser breve, debe abarcar todas las cuestiones pertinentes, y debe explorar en detalle aquellas cuestiones que se consideran de particular interés para el estudio. Claramente, hay que realizar compromisos: primero, entre la reducción de un cuestionario que se considera que es demasiado largo, pero conservando los ítems suficientes para proporcionar una cobertura completa de la CVRS (validez de contenido); y, en segundo lugar, entre el mantener toda la amplitud de cobertura de los ítems con el objetivo de realizar la evaluación detallada y en profundidad de aspectos específicos.

Los esfuerzos para desarrollar cuestionarios cortos se han centrado principalmente en la reducción de instrumentos existentes, y los métodos más frecuentemente utilizados en la reducción de cuestionarios se basan en la

estadística. Al igual que ocurre en el proceso de validación de instrumentos de CVRS, en el proceso de reducción de cuestionarios el método más utilizado ha sido el AF, método clasificado dentro de la psicometría clásica. Sin embargo, la literatura más actual también propone para la reducción de cuestionarios la utilización de procedimientos más modernos basados en la TRI (Prieto y cols., 2003).

1.8. La medición de la calidad de vida relacionada con la salud en enfermedades crónicas

La CVRS tiene un lugar destacado en la investigación de resultados de salud (Clancy y Eisenberg, 1998; Epstein y Sherwood, 1996) como un parámetro de medición de los estados de salud y evaluación de los resultados de los cuidados médicos, en especial en el campo de enfermedades crónicas. En Estados Unidos la *Food and Drug Administration* obliga a incluir la CVRS como un resultado para la evaluación de los nuevos tratamientos aplicables a las enfermedades crónicas. En Europa, la Agencia Europea del Medicamento no obliga aún, pero si recomienda la inclusión de esta medida en la evaluación de nuevos tratamientos contra el cáncer.

Kaplan (2003) en un artículo sobre la utilidad de los indicadores de calidad de vida en el sistema de salud, describe las dos posturas vigentes dentro de la medicina: el Modelo Biomédico tradicional, y el Modelo de Resultados. El Modelo Biomédico tradicional está orientado a la detección y cura de las enfermedades, dando prioridad a los resultados que provienen de pruebas médicas, la esperanza de vida lograda, la proporción de mortalidad detectada y el número de días de incapacidad. Por otro lado, dentro del Modelo de Resultados, la meta más importante de la medicina es que la gente viva lo mejor posible durante el mayor número de años. Pero en algunas ocasiones, el diagnóstico y el tratamiento de una enfermedad no conducen a un aumento del estado de salud, como puede ser por ejemplo el caso de las enfermedades crónicas. En este caso, según el Modelo Biomédico tradicional no puede darse un éxito ya que la cura no existe. En este tipo de enfermedades, el

Modelo de Resultados considera que lo importante es conocer la percepción del paciente acerca de su condición. Se centra en evaluar la adaptación del paciente a su enfermedad y evolución de las características sociales y psicológicas del paciente que son componentes básicos de la medición de la CVRS.

Por lo tanto, los instrumentos de evaluación de la CVRS suponen una herramienta de gran utilidad dentro del campo de las enfermedades crónicas, y es por ello que para el desarrollo de este trabajo hemos escogido diferentes estudios con muestras de pacientes crónicos que describiremos a continuación. Además, hemos seleccionado unas enfermedades crónicas con mayor afectación física y otras con mayor afectación mental. Estos grupos de pacientes son: pacientes con obesidad mórbida, pacientes con artrosis de cadera, y pacientes con trastorno de la conducta de la alimentación (TCA).

1.8.1. Obesidad mórbida

La obesidad es una enfermedad crónica que se caracteriza por un exceso de grasa en el organismo y que se manifiesta por un aumento del peso corporal. Para la valoración del grado de obesidad se utiliza el Índice de Masa Corporal (IMC), y se define como obesidad un IMC > 30 kg/m^2, y obesidad mórbida un IMC > 40 kg/m^2.

La Sociedad Española para el Estudio de la Obesidad publicó en el año 2003 (Aranceta y cols., 2003) un estudio en el que mostraba que la prevalencia de la obesidad para el conjunto de la población española entre 25 y 60 años era de 14,5% (13,39% en varones y 15,75% en mujeres), y el de la obesidad mórbida era de 0,5% (0,3% en varones y 0,7% en mujeres). En la Comunidad Autónoma del País Vasco se realizó una Encuesta de Salud (Anitua y cols., 1997) siendo la prevalencia de obesidad mórbida, para el grupo de 16 a 64 años del 0,25%. En un estudio prospectivo a 10 años realizado en la Comunidad Autónoma del País Vasco sobre la incidencia y factores de riesgo de la Diabetes Mellitus II se registraron los datos de peso y talla. Al comparar los datos de 1985 y 1995 la prevalencia de la

obesidad mórbida se ha observado que ha duplicado, pasando de 0,2% a 0,4% (Vázquez y cols., 2000).

La obesidad mórbida es una enfermedad con una gran repercusión en el paciente en términos de pérdida de esperanza de vida y de calidad de vida. El impacto de esta enfermedad crónica sobre la esperanza de vida viene derivada del exceso de morbimortalidad asociado. Existen una serie de comorbilidades que aparecen o se complican debido a la obesidad mórbida y que mejoran o se curan con la pérdida sustancial de peso. Entre ellas se encuentran la enfermedad coronaria, la hipertensión arterial, la diabetes mellitus, el síndrome de apnea del sueño, la osteoartrosis grave en articulaciones de carga, así como otras menores como varices, colelitiasis, infertilidad, reflujo gastro-esofágico y alteraciones menstruales.

La obesidad mórbida también tiene un gran impacto sobre la CVRS, ya que se asocia con alteraciones socioeconómicas y psicosociales, así como con depresión y pérdida de la autoestima. En un test de calidad de vida (Fontaine y Bartlett, 1998) en estos pacientes, se observa que los aspectos más afectados son la función física, las limitaciones por problemas físicos, el dolor corporal, la disminución en la percepción de la salud general, la vitalidad, la función social, y las limitaciones del rol por problemas emocionales y de salud mental. El impacto sobre la autoestima y la función sexual es mayor en mujeres que en hombres (Kolotkin y cols., 1997).

La obesidad mórbida es con frecuencia refractaria al tratamiento médico convencional (medidas dietéticas, ejercicio físico y terapias de modificación de conducta). La cirugía bariátrica es una opción en los pacientes en los que ha fracasado el tratamiento convencional. La cirugía bariátrica se realiza para conseguir una pérdida efectiva de peso patológico y un mantenimiento de dicha pérdida a largo plazo, así como corregir o controlar la patología asociada a la obesidad mórbida y mejorar la calidad de vida del paciente, con un mínimo número de complicaciones.

Por tanto, teniendo en cuenta la considerable disminución de la calidad de vida que conlleva esta enfermedad, la medición de la CVRS es fundamental para medir los resultados del tratamiento en pacientes con obesidad mórbida siendo un indicador particularmente apropiado del impacto que la enfermedad tiene sobre el paciente. Debido a la importancia que el impacto de la obesidad mórbida tiene sobre aspectos de la CVRS tales como el funcionamiento psicológico y físico, hemos elegido para su medición el cuestionario especifico Obesity-related Problems scale (OP) (Karlsson y cols., 2003; Sullivan y cols., 1993). Este es un cuestionario diseñado para medir el impacto de la obesidad en el funcionamiento psicosocial, que consta de ocho preguntas cubriendo un único dominio (Anexo I), cuyo diagrama de flujo se muestra en la Figura 1.3.

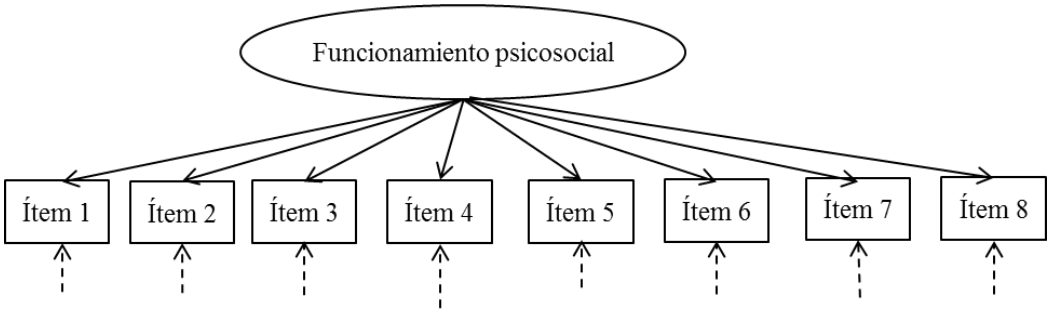


Figura 1.3. Diagrama de flujo que representa la estructura interna del cuestionario Obesity-related Problems scale (OP).

1.8.2. Artrosis de cadera

La artrosis es una enfermedad crónica y una causa importante y frecuente de dolor e incapacidad en los paises desarrollados, particularmente en las poblaciones que envejecen (Mahon y cols., 2002; Núñez y cols., 2009; Sociedad Española de Reumatología, 2001), suponiendo así un grave problema de salud pública (Cushnaghan y cols., 2007; Rat y cols., 2010). La artrosis es más frecuente en las articulaciones de cadera o de rodilla, causando una incapacidad considerable y el aislamiento social en personas de edad avanzada, con la consiguiente pérdida de CVRS (Núñez y cols., 2009; Quintana y cols., 2008a). Aproximadamente entre el 7%

y el 11% de la población de los países desarrollados padecen artrosis clínica, y del 27% al 44% tienen lesiones radiológicas (Felson y cols., 1995; Felson y Zhang, 1998). Según un estudio realizado en el País Vasco, la prevalencia de artrosis de cadera es del 7,4% (Quintana y cols., 2008b), aumentando esta con la edad. Varios estudios, han publicado la prevalencia de dolor de cadera en la población adulta, pero con resultados diversos (Dawson y cols., 2004; Quintana y cols., 2008a). Sin embargo, a pesar del rango de edad de los pacientes en esos estudios, todos coinciden en que la prevalencia de la artrosis de cadera es mayor entre aquellos que son mayores de 65 años. Al incrementarse la prevalencia de la artrosis con la edad, y el envejecimiento de la población de muchos países, como el nuestro, llevará a convertir la artrosis en un problema cada vez más importante de salud.

La artroplastia total de cadera (ATC) es un procedimiento que se utiliza principalmente en pacientes con un diagnóstico de artrosis, y que se lleva a cabo cada vez más en los países desarrollados (Frankel y cols., 1999; Knutson y cols., 1994). La ATC se considera que es una intervención coste-efectiva que mejora la calidad de vida del paciente, reduce el dolor y aumenta su capacidad funcional asociada a la artrosis (Cushnaghan y cols., 2007; Escobar y cols., 2007a y 2007b; Rat y cols., 2010; Quintana y cols., 2005). Si bien la ATC está entre los procedimientos más efectivos en términos de beneficios para el paciente, también está entre los procedimientos que comportan un coste económico elevado (Quintana y cols., 2006). Por tanto, las limitaciones en la actividad de esta población cada vez mayor de personas de edad avanzada deben ser evaluadas para predecir las futuras políticas de salud pública. Los datos sobre el nivel de incapacidad y evolución en el tiempo de este aumento de la población son necesarios para la planificación de los servicios de salud y los recursos presupuestarios y para la mejora en informar a los pacientes acerca de sus posibles dificultades después de la cirugía.

Aunque la mayoría de los informes de evaluación de la efectividad de la ATC se han centrado en los aspectos técnico-quirúrgicos, existe un creciente interés en el punto de vista del paciente mediante el uso de cuestionarios auto-administrados

(Kaufman, 2001). La medición del estado de salud en la artrosis ha sufrido una evolución progresiva en los últimos 60 años, con cambios más rápidos en los últimos 25 años (Escobar y cols., 2007b). Teniendo en cuenta que las principales razones para llevar a cabo una intervención de esta naturaleza son dolor en las articulaciones, limitaciones funcionales o ambas cosas, además de evaluar objetivamente los resultados de las intervenciones mediante valoraciones clínicas clásicas, también se pueden utilizar medidas subjetivas obtenidas de los pacientes. La medición de la CVRS ha sido cada vez más reconocida como un medio para medir los resultados de la ATC, siendo este un indicador importante del impacto en la población de medidas terapéuticas en este tipo de enfermedades. Es particularmente apropiado medir la CVRS en pacientes con artrosis debido a que el debilitamiento crónico de esta enfermedad parece conllevar una considerable disminución en la calidad de vida. Parece por tanto ser un buen indicador de los efectos globales de la artrosis en la vida del paciente y de los efectos del tratamiento.

Entre los diferentes instrumentos de medida de CVRS en pacientes con artrosis de cadera, nos encontramos con el cuestionario especifico Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) (Bellamy y cols., 1988a y 1988b). Este es uno de los instrumentos más utilizados para medir sintomatología y función en pacientes con artrosis de cadera o rodilla (Anderson y cols., 1996; Hawker y cols., 1995 y 1998). Este cuestionario fue desarrollado para evaluar cambios clínicamente importantes y relevantes para el paciente en el estado de salud tras un tratamiento (Bellamy, 2000). El WOMAC ha sido traducido y validado en población española (Escobar y cols., 2002; Escobar y cols., 2007b; Quintana y cols., 2005). El cuestionario consta de 24 preguntas y cubre tres dominios: dolor, capacidad funcional y rigidez. La Figura 1.4 muestra la estructura interna del WOMAC y el cuestionario completo se incluye en el Anexo II.

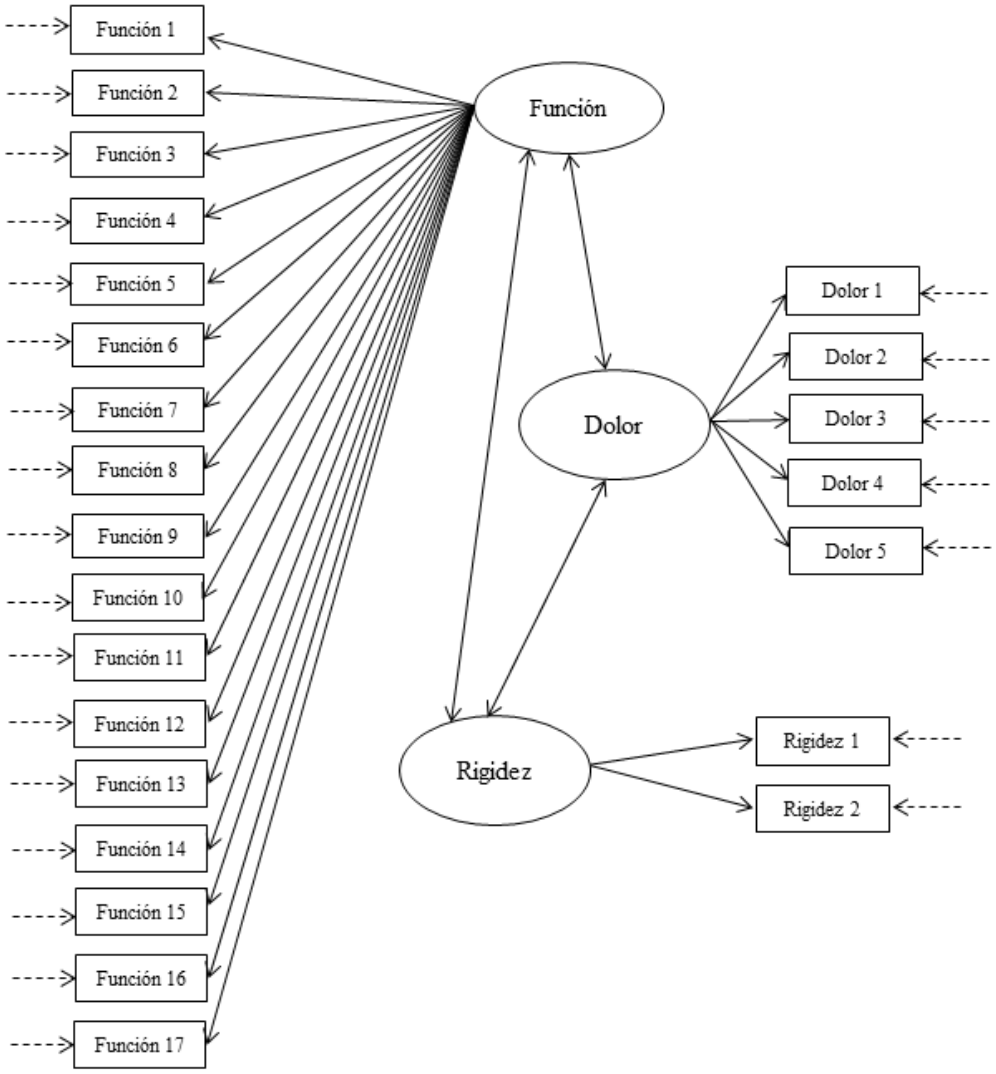


Figura 1.4. Diagrama de flujo que representa la estructura interna del cuestionario Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC).

1.8.3. Trastornos de la conducta de la alimentación

Los TCA son un problema relativamente común entre personas jóvenes y adolescentes. Los diferentes subtipos dentro de los TCA incluyen la Anorexia Nerviosa (AN), la Bulimia Nerviosa (BN), el Trastorno por Atracón y el Trastorno de la Alimentación No Específico (TANE). Estos subtipos se distinguen entre sí por conllevar conductas diferentes, pero todos ellos comparten que la persona afectada tiene un problema a la hora de alimentarse adecuadamente, una

insatisfacción corporal muy alta, que puede llegar a poner en peligro su vida, y una autoestima muy baja. Si además añadimos que estos trastornos ocurren más frecuentemente en personas jóvenes (desde los 15 a los 24 años) la magnitud del problema aumenta, ya que de su recuperación total o parcial va a depender su futuro.

Datos epidemiológicos (Hoek, 2006) señalan que las tasas de incidencia de AN son de 8 nuevos casos cada 100.000 habitantes por año, y de 13 nuevos casos cada 100.000 habitantes por año en la BN. Hay que señalar que aunque los estudios hayan sido muy rigurosos a la hora de calcular estas tasas de incidencia, muchos casos de AN y BN en la población no han sido diagnosticados aún, ya que las propias personas enfermas tienen la tendencia a esconder su trastorno a los demás. Con respecto a la AN la prevalencia es de 0,3% y en la BN es del 1% dentro del grupo de las mujeres jóvenes. La tasa de prevalencia del Trastorno por Atracón en los Estados Unidos es del 2,6% de casos entre las mujeres de raza blanca de 18 a 40 años, y entre las mujeres de raza negra del mismo rango de edad la tasa de prevalencia ascendía al 4,5% (Hoek y van Hoeken, 2003). Finalmente, el 60% de los diagnósticos de TCA que se realizan en consultas externas son de TANE, por lo que son mayoría, ya que el 14,5% son de AN y el 25,5% de BN (Fairburn y Bohn, 2005).

Como ya hemos mencionado, la CVRS se puede definir como el impacto que la enfermedad y su tratamiento tienen en las diferentes áreas del individuo, siendo éstas principalmente tres, el área física, el área psíquica y el área social (Revicki y cols., 2000). Las personas con un TCA pueden tener afectadas estas áreas de diferentes formas. Hay diversos estudios que han valorado el impacto del TCA en cada una de estas áreas (Miller y cols., 2005), mostrando como los TCA afectan negativamente a los diversos ámbitos de la calidad de vida de las personas que lo padecen. Por todo ello, es fundamental la evaluación de la CVRS en este tipo de pacientes. Aunque existen numerosos estudios en los que han medido la CVRS en estos pacientes, la mayoría de los estudios han utilizado cuestionarios genéricos para medir el impacto que la enfermedad tiene en aspectos físicos, mentales y

sociales (Hay y Mond, 2005). Los primeros cuestionarios específicos para personas afectadas de un trastorno de la alimentación fueron publicados simultáneamente entre los años 2006 y 2007 (Abraham y cols., 2006; Adair y cols., 2007; Engel y cols., 2006; Las Hayas y cols., 2006 y 2007). Uno de estos cuestionarios es el Health Related Quality of Life for Eating Disorders version 2 (HeRQoLEDv2) (Las Hayas y cols., 2006 y 2007), que contiene 55 preguntas y cubre nueve dominios: síntomas, conductas restrictivas, imagen corporal, salud mental, rol emocional, rol físico, rasgos de personalidad, relaciones sociales y atracones. La Figura 1.5 muestra el diagrama de flujo de las relaciones y estructura interna del cuestionario HeRQoLEDv2 y el cuestionario aperece en su totalidad en el Anexo III.

1.9.Justificación

Como ya hemos mencionado, la medición de la CVRS es de gran importancia y de gran utilidad dentro del ámbito sanitario, y más concretamente dentro de las enfermedades crónicas. La manera de medir la CVRS es a través de cuestionarios, y para determinar la aceptabilidad de un cuestionario de medición de la CVRS es necesario que este cumpla con ciertos atributos, entre los que se encuentran las propiedades psicométricas, tales como la fiabilidad, la validez y la sensibilidad al cambio.

De entre las diferentes propiedades psicométricas, la validez de constructo es el aspecto más difícil y más controvertido, y a su vez, es el más susceptible de exploración mediante el análisis numérico y estadístico. Además, dentro de la validez de constructo podemos decir que la validez de estructura es el que juega un papel más importante, siendo también el tema con mayor controversia. Los métodos estadísticos a utilizar para la evaluación de la validez convergente y discriminante, o la validez de grupos conocidos, o incluso la validez de criterio o la fiabilidad, así como la sensibilidad al cambio, son métodos bien establecidos y ampliamente aceptados. Sin embargo, no ocurre lo mismo con los métodos estadísticos a utilizar para la evaluación de la validez de estructura. Como ya hemos visto, existen diferentes perspectivas y teorías de enfoque de la validez de

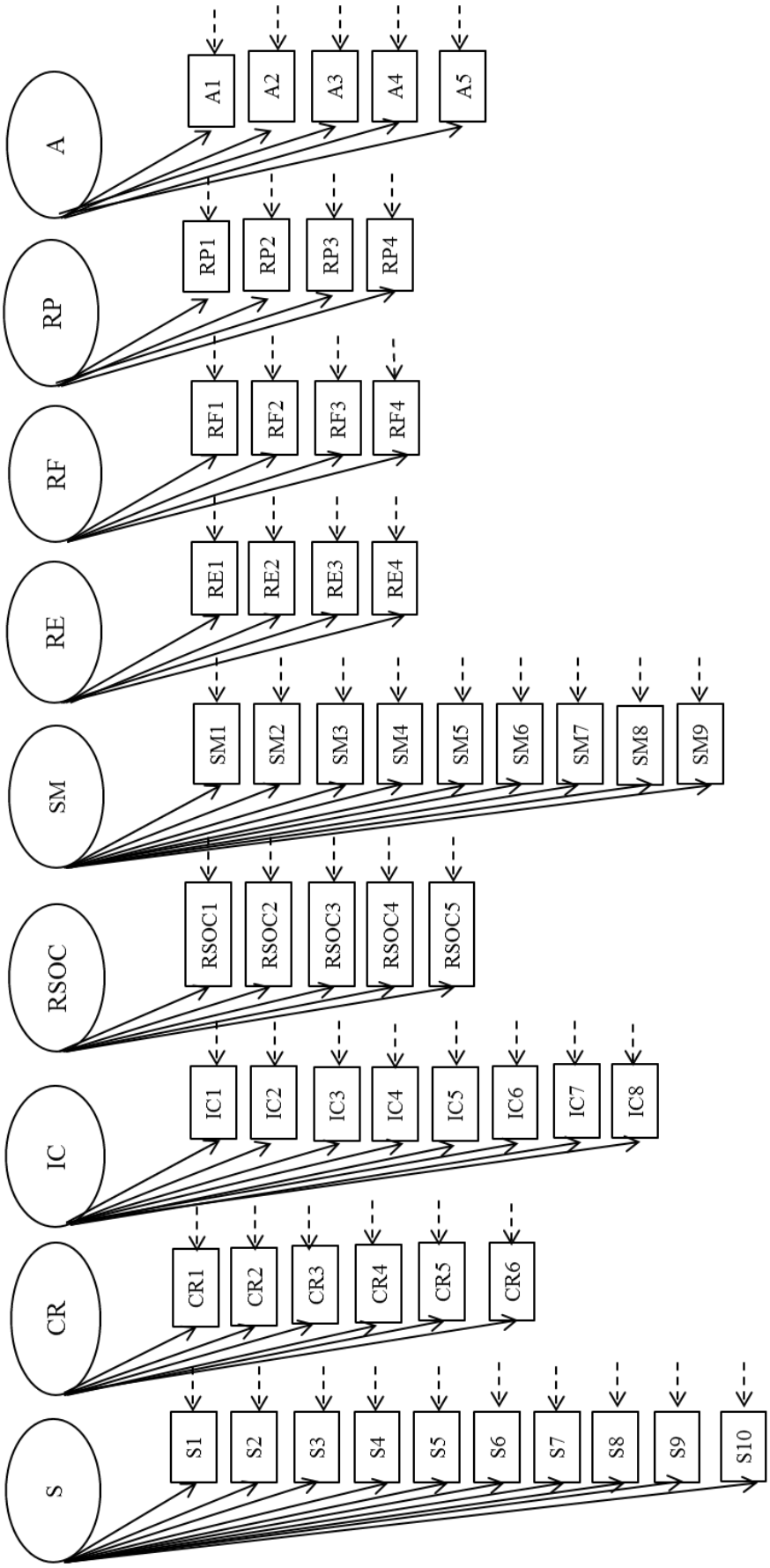


Figura 1.5. Diagrama de flujo que representa la estructura interna del cuestionario Health Related Quality of Life for Eating Disorders version 2 (HeRQoLEDv2). S: Síntomas; CR: Conductas restrictivas; IC: Imagen corporal; RSOC: Relaciones sociales; SM: Salud mental; RE: Rol emocional; RF: Rol físico; RP: rasgos de personalidad; A: Atracones.

estructura, desde la TCT hasta la TRI, e incluso dentro de cada una de las teorías, existen diferentes métodos.

Como ya hemos comentado, aunque la literatura más reciente recomienda la utilización de los métodos basados en la TRI, su utilización en el campo de la medición de la CVRS no es aún muy extendida. Una búsqueda actualizada hasta la fecha de diciembre del 2015 en la base de datos de *Medline* con la palabra clave "factor analysis" nos muestra 37.230 artículos. Si la palabra clave es "item response theory" el número de artículos se reduce a 1.520. Este número sigue disminuyendo a 1.143 artículos para la palabra clave "rasch analysis", a 115 para "rating scale model" y a únicamente 80 artículos en caso de considerar "graded response model" como palabra clave.

El abanico de métodos disponibles para la evaluación de la validez de estructura y reducción de cuestionarios es amplio. El investigador debe decidir el método a utilizar para validar y/o reducir un cuestionario, pero no existe un consenso sobre cual utilizar y cuando. Esto hace que el estudio de la validez de estructura sea el aspecto más complejo y de mayor controversia dentro del proceso de validación o reducción de un instrumento. Dependiendo de las características del cuestionario y del estudio unas técnicas pueden aportar mayores beneficios que otras. Por tanto, conocer las ventajas y desventajas de cada una de las técnicas es una cuestión relevante, para así poder aplicarlas de forma adecuada. Más aún, la combinación de varias técnicas podría proporcionar mayor evidencia de la validez de un instrumento. La literatura actual sobre las técnicas de validación y reducción de cuestionarios más adecuadas en cada contexto, las ventajas que aportan las técnicas modernas sobre las clásicas o sobre los beneficios de la combinación de técnicas no es ni suficiente ni clarificadora.

Por todo ello, es necesario establecer las ventajas y desventajas de cada una de las diferentes técnicas dependiendo de los datos disponibles y del tipo de cuestionario de CVRS, para así saber cuándo aplicar una técnica u otra. Así, es en este aspecto donde centraremos sobre todo el trabajo que aquí se presenta, mediante la

aplicación de diferentes técnicas estadísticas tanto de la psicometría clásica como de la moderna a diferentes cuestionarios de CVRS. Aunque en el Capítulo 3 presentaremos la metodología para el análisis de todas las propiedades psicométricas planteadas, fiabilidad, validez y sensibilidad al cambio, el objetivo del presente trabajo se centra principalmente en la aplicación de diferentes técnicas estadísticas para el análisis de la validez de estructura y reducción de cuestionarios de CVRS. En cuanto a las técnicas estadísticas, hemos seleccionado tanto técnicas estadísticas dentro de la denominada TCT como de la TRI. Así, dentro de la TCT aplicaremos desde el AFE, el AFC y los MEE más tradicionales, hasta los métodos de AFC y MEE para datos categóricos. Por otro lado, dentro de los métodos de la TRI, aplicaremos la generalización tanto de los modelos de 1–parámetro como de 2–parámetros para datos politómicos. Concretamente aplicaremos los métodos de Rating Scale Model y Graded Response Model.

Para la aplicación de estos métodos hemos seleccionado diferentes muestras de pacientes, todos ellos pacientes con enfermedades crónicas tales como obesidad mórbida, artrosis de cadera y TCA. Por todo ello, en la presente tesis se describen las diferentes técnicas estadísticas para así compararlas aplicándolas a diversos cuestionarios utilizados en las diferentes cohortes de pacientes. Concretamente, en relación a los pacientes con obesidad mórbida, trabajaremos con el cuestionario específico OP scale, cuyo diagrama de flujo se muestra en la Figura 1.3, realizando la traducción, adaptación y validación del cuestionario al español. Se estudiarán tanto la fiabilidad como aspectos de la validez, tales como la validez convergente, la validez de grupos conocidos, así como la validez de estructura. Para este último se emplearán técnicas dentro de la TCT como el AFE y AFC. Estos resultados se presentarán en el Capítulo 4 de la presente tesis. Concretamente, este trabajo fue uno de mis primeros contactos con el campo de la validación de cuestionarios de CVRS, y es el que me ha servido como base para seguir avanzando en este campo. Este trabajo es un ejemplo clásico de la validación de un instrumento, y el que generó mi interés en esta línea de investigación.

En cuanto a los pacientes con artrosis de cadera, trabajaremos con el cuestionario específico WOMAC, cuyo diagrama de flujo se muestra en la Figura 1.4. Realizaremos una propuesta de una versión reducida de dicho cuestionario, basándonos en versiones ya reducidas, y su posterior validación. La Figura 1.6 muestra el diagrama de flujo de la versión reducida y el cuestionario reducido se incluye en el Anexo IV. Se estudiará la fiabilidad de la versión reducida, así como la validez convergente y discriminante, la validez de grupos conocidos, la validez concurrente, y la validez de estructura. En este caso, para la validez de estructura se empleará el AFC dentro de la TCT, así como el Rating Scale Model dentro de la TRI. Además, también se estudiará la sensibilidad al cambio de la versión reducida. Los resultados de este estudio se presentarán en el Capítulo 5 de esta tesis.

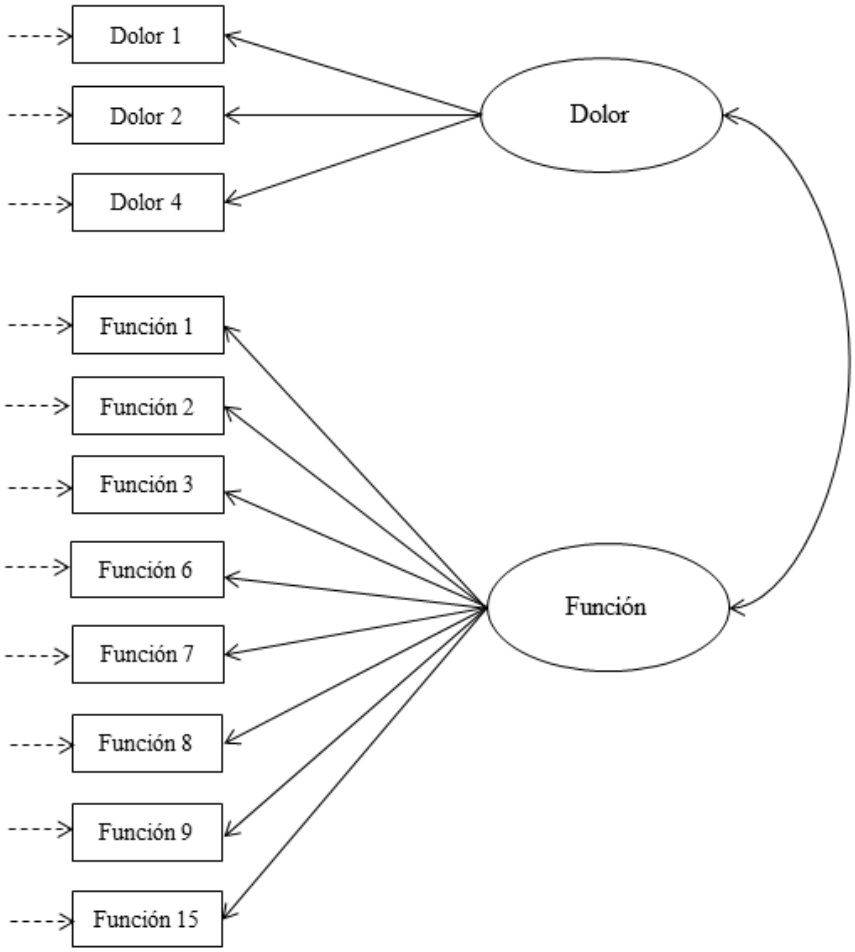


Figura 1.6. Diagrama de flujo que representa la estructura interna de la versión reducida del cuestionario Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC). Los ítems están nombrados tal y como aparecen en la versión original.

Por último, en relación a los pacientes con trastorno de la alimentación, trabajaremos con el cuestionario HeRQoLEDv2, cuyo diagrama de flujo se muestra en la Figura 1.5. En este caso, en primer lugar se confirmará la existencia de una estructura interna de segundo orden con dos medidas resumen, tal y como se muestra en la Figura 1.7. En segundo lugar, se obtendrá una versión reducida del cuestionario, mediante la aplicación de técnicas tanto basadas en la TCT como en la TRI. Comenzaremos aplicando los MEE para estudiar la estructura subyacente, y continuaremos aplicando el Rating Scale Model para la reducción del cuestionario. Estos resultados se presentarán en el Capítulo 6 de la presente tesis. Finalmente, una vez obtenida la versión reducida del cuestionario, denominada Health Related Quality of Life for Eating Disorders Questionnaire-Short Version (HeRQoLED-S) (Anexo V) y cuyo diagrama de flujo se muestra en la (Figura 1.8), se validará esta versión reducida en otra muestra de pacientes con TCA. En este caso, para la validez de estructura se emplearán los MEE para datos categóricos dentro de la TCT, así como el Graded Response Model dentro de la TRI. Además, se estudiará la fiabilidad, validez convergente y discriminante, validez de grupos conocidos y la sensibilidad al cambio de esta versión reducida. Así, en el Capítulo 7 se presentarán los resultados obtenidos de esta validación del HeRQoLED-S.

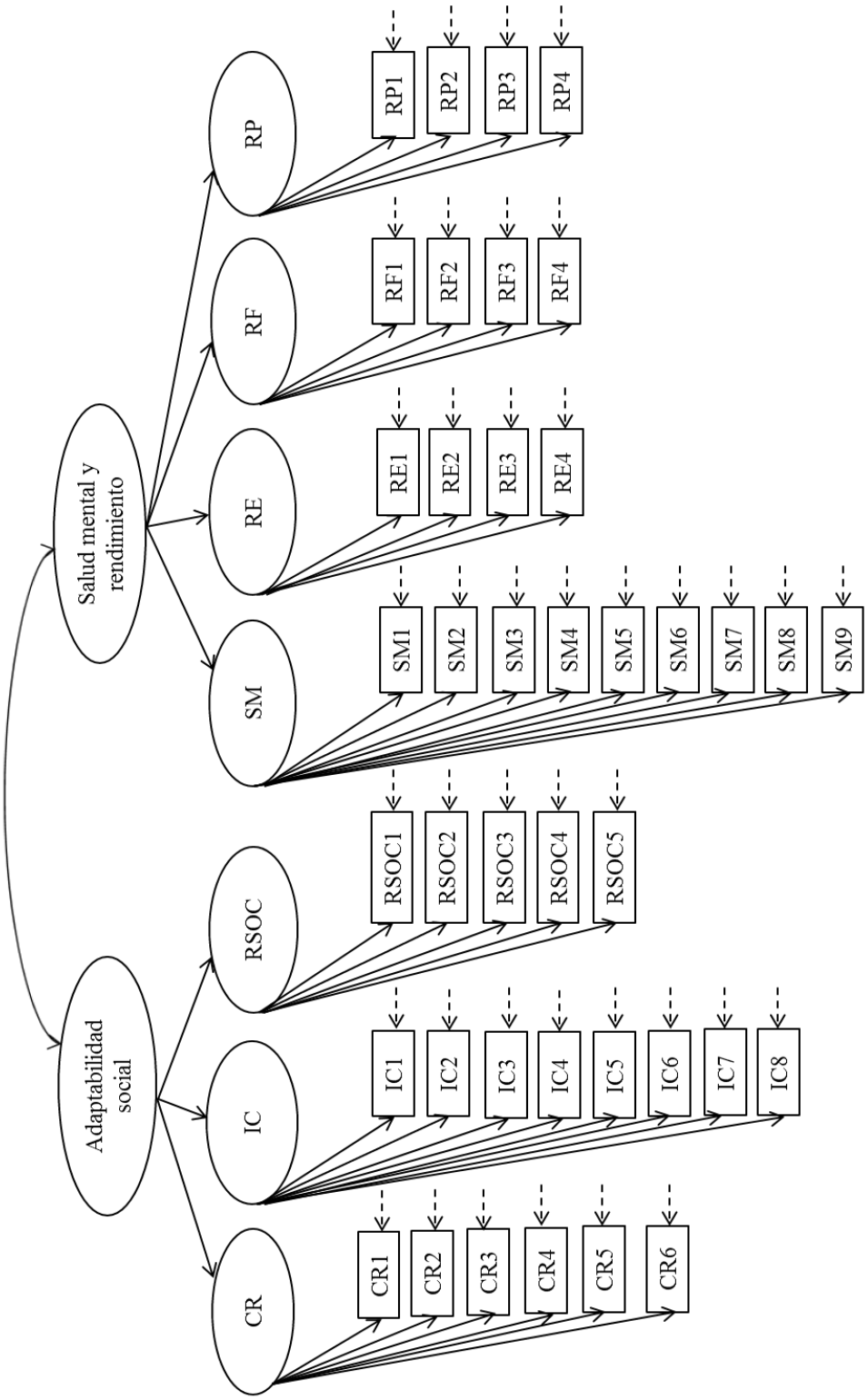


Figura 1.7. Diagrama de flujo que representa la estructura interna de segundo orden del cuestionario Health Related Quality of Life for Eating Disorders version 2 (HeRQoLEDv2). CR: Conductas restrictivas; IC: Imagen corporal; RSOC: Relaciones sociales; SM: Salud mental; RE: Rol emocional; RF: Rol físico; RP: rasgos de personalidad.

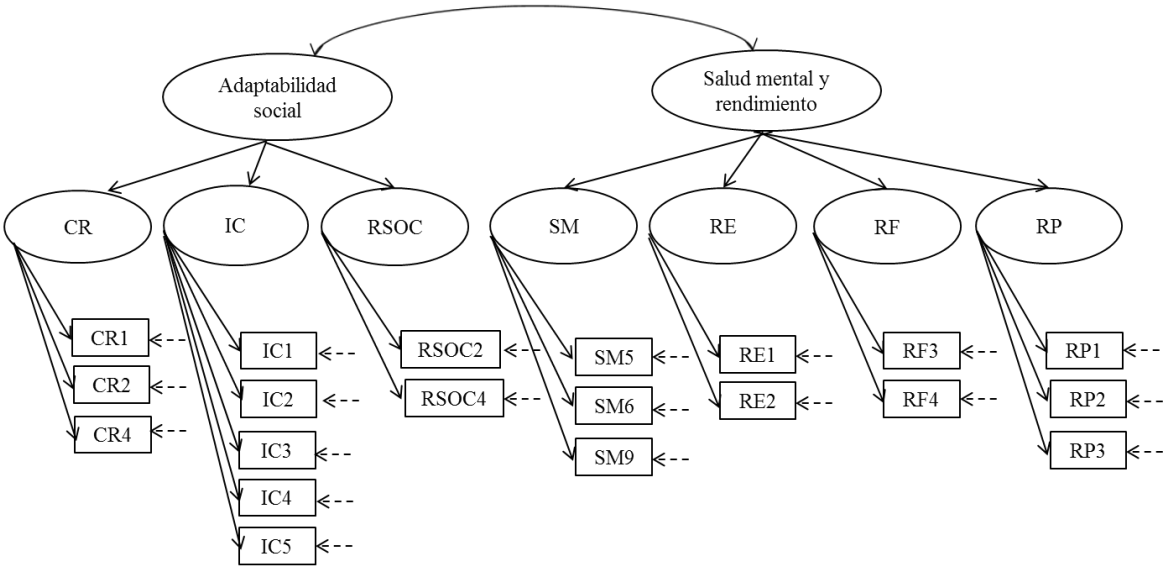


Figura 1.8. Diagrama de flujo que representa la estructura interna de segundo orden del cuestionario reducido Health Related Quality of Life for Eating Disorders-Short Version (HeRQoLED-S). CR: Conductas restrictivas; IC: Imagen corporal; RSOC: Relaciones sociales; SM: Salud mental; RE: Rol emocional; RF: Rol físico; RP: rasgos de personalidad.

Capítulo 2

Hipótesis y objetivos

2.1. Hipótesis

Hipótesis generales

1. Los métodos estadísticos basados en la teoría clásica del test (TCT) y los basados en la teoría de la respuesta al ítem (TRI) son complementarios entre sí, proporcionando cada uno de ellos diferente tipo de soporte en el proceso de validación o reducción de cuestionarios de calidad de vida relacionada con la salud (CVRS).

2. La combinación de diferentes métodos estadísticos, tanto de la TCT como de la TRI, en el proceso de validación y reducción de cuestionarios de CVRS, proporciona evidencia de su validez y de sus buenas propiedades

psicométricas, proporcionando así versiones científicamente validadas en castellano, para diferentes ámbitos de la salud.

Hipótesis específicas

1. Los métodos estadísticos de la TCT permiten comprobar la fiabilidad y validez de la versión española del cuestionario Obesity-related Problems scale (OP) diseñado para medir el impacto de la obesidad mórbida en el funcionamiento psicosocial.

2. La combinación de métodos estadísticos de la TCT así como de la TRI, proporcionan mayor evidencia sobre la validez y sobre las buenas propiedades psicométricas de una versión reducida del cuestionario Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), empleado para la medición de sintomatología y función en pacientes con artrosis de extremidad inferior.

3. El cuestionario Health Related Quality of Life for Eating Disorders version 2 (HeRQoLEDv2) diseñado para medir la CVRS en pacientes con un trastorno de la conducta de la alimentación, tiene una estructura interna subyacente de segundo orden, con dos medidas resumen; y los métodos estadísticos de la TCT permiten confirmar la existencia de dicha estructura de segundo orden.

4. La combinación de métodos estadísticos de la TCT así como de la TRI, permiten reducir el cuestionario HeRQoLEDv2 y estudiar su fiabilidad, validez y sensibilidad al cambio, proporcionando mayor evidencia sobre sus buenas propiedades psicométricas.

2.2. Objetivos

Objetivos generales

1. Validar y/o reducir cuestionarios de CVRS mediante la combinación de métodos estadísticos, tanto de la perspectiva de la TCT como de la TRI, y evaluar la complementariedad de dichos métodos.

2. Proporcionar herramientas de medición de la CVRS científicamente validadas y con buenas propiedades psicométricas en castellano para diferentes ámbitos de la salud.

Objetivos específicos

1. Traducir y adaptar al español el cuestionario OP, y estudiar su fiabilidad y validez mediante diferentes métodos estadísticos de la TCT.

2. Proponer una versión reducida del cuestionario WOMAC y estudiar la fiabilidad, validez y sensibilidad al cambio de dicha versión reducida mediante la combinación de métodos estadísticos de la TCT y de la TRI.

3. Confirmar la existencia de una estructura de segundo orden con dos medidas resumen en el cuestionario HeRQoLEDv2 mediante métodos estadísticos de la TCT.

4. Reducir el cuestionario HeRQoLEDv2 y estudiar la fiabilidad, validez y sensibilidad al cambio de dicha reducción mediante diferentes métodos estadísticos de la TCT así como de la TRI, y proporcionar así una versión reducida científicamente validada en castellano.

Capítulo 3

Metodología

Los resultados de esta tesis doctoral están publicados como artículos originales en revistas científicas internacionales, y estos serán presentados en los Capítulos 4, 5, 6 y 7 de la presente tesis. Por ello, por no repetirnos, en este capítulo de metodología común describiremos los métodos utilizados para la validación y reducción de cuestionarios de calidad de vida relacionada con la salud muy brevemente, ya que estos se presentan con mayor detalle en los capítulos dedicados a los resultados.

La parte metodológica de este capítulo consta de una primera parte de diseño del estudio, ámbito, sujetos, materiales y recogida de datos, diferente para cada patología seleccionada, y de una segunda parte de análisis estadístico para el estudio de validación y reducción de cuestionarios. Esta segunda parte es la que describiremos de manera global para todas las patologías y cuestionarios, aunque en cada caso se aplican unos métodos u otros.

3.1. Diseño, ámbito, sujetos, materiales y recogida de datos

3.1.1. Estudio de obesidad mórbida

La metodología empleada para este estudio se presentará con más detalle en el primer capítulo dedicado a resultados, el Capítulo 4.

Diseño, ámbito y sujetos

Se realizó un estudio de cohortes prospectivo de pacientes con obesidad mórbida que se encontraban en lista de espera para ser intervenidos de cirugía bariátrica, en dos hospitales del País Vasco, Hospital Universitario Araba–Sede Txagorritxu y Hospital Universitario Donostia, pertenecientes a Osakidetza. El periodo de reclutamiento fue de mayo del 2005 a mayo del 2006.

Materiales y recogida de datos

Los pacientes cumplimentaron el cuestionario específico Obesity-related Problems scale (OP) (Karlsson y cols., 2003; Sullivan y cols., 1993), así como los cuestionarios genéricos Short Form 36 (SF-36) (Ware y Sherbourne, 1992) y EQ-5D-3L (EuroQoL Group, 1990). Además, se recogieron datos sociodemográficos de los pacientes y datos clínicos mediante la revisión de las historias clínicas.

Proceso de traducción del cuestionario Obesity-Related Problem Scale

El cuestionario OP fue adaptado y traducido al español siguiendo el método de traducción-retrotraducción (Aaronson y cols., 1992). El cuestionario fue traducido de su idioma original (inglés) al español, de manera independiente, por dos personas bilingües, cuya lengua nativa era el español pero con un nivel de inglés competente. A ambos traductores se les indicó que las traducciones deberían ser semánticas y no literales, buscando la equivalencia conceptual e idiomática. El equipo investigador comparó ambas traducciones, y las diferencias se resolvieron

llegando a un consenso entre los dos traductores, para así desarrollar la primera versión del cuestionario adaptado.

A continuación se llevó a cabo la retrotraducción al inglés de esta primera versión. Esta fue traducida del español al inglés de manera independiente por dos traductores cuya lengua nativa era inglesa pero que tenían un nivel muy fluido de español. Estos traductores eran "ciegos" al cuestionario original. Ambas traducciones fueron comparadas entre sí y con la versión original del cuestionario por el equipo investigador, así como por los dos traductores, y fueron enviadas al autor del cuestionario original. Tras discutir las diferencias encontradas, se elaboró la segunda versión del cuestionario.

Esta segunda versión española del cuestionario fue probada en una pequeña muestra de 20 pacientes con obesidad mórbida, que estaban en lista de espera para ser intervenidos de cirugía bariátrica. El objetivo de este estudio piloto era comprobar que esta versión española del cuestionario no contenía preguntas confusas ni de difícil comprensión. Se les preguntó a los pacientes si alguna de las preguntas del cuestionario eran difíciles de entender, difíciles de responder, o confusas o si se podían enunciar de alguna manera diferente. A partir de estos resultados se elaboró la versión final del cuestionario, que fue enviada al autor del cuestionario original.

3.1.2. Estudio de artrosis de cadera

La metodología empleada para este estudio se presentará con más detalle en Capítulo 5 de la presente tesis.

Diseño, ámbito y sujetos

Este estudio incluye datos de dos cohortes prospectivas independientes. Se reclutaron pacientes con artrosis de cadera que se encontraban en lista de espera para ser intervenidos de artroplastia total de cadera en diferentes hospitales

pertenecientes a Osakidetza. Una de las cohortes fue reclutada entre marzo de 1999 y marzo de 2000 (cohorte 1), mientras que la otra se reclutó entre septiembre de 2003 y septiembre de 2004 (cohorte 2).

Materiales y recogida de datos

Los pacientes cumplimentaron el cuestionario específico Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) (Bellamy y cols., 1988b), el cuestionario genérico SF-36 (Ware y Sherbourne, 1992), así como otras preguntas adicionales en relación al grado de dolor y función, mientras se encontraban en lista de espera para ser intervenidos. Seis meses después de la intervención, los pacientes volvieron a cumplimentar estos mismos cuestionarios. Además, se recogieron datos sociodemográficos y datos clínicos mediante la revisión de las historias clínicas, tanto antes de la intervención como en el seguimiento a los seis meses de la cirugía.

3.1.3. Estudio de trastornos de la conducta de la alimentación

La metodología empleada para este estudio se presentará con más detalle en los Capítulos 6 y 7.

Diseño, ámbito y sujetos

Este estudio incluye datos de dos cohortes prospectivas. En ambos casos se reclutaron pacientes con diagnóstico de trastorno de la conducta de la alimentación que estaban siendo tratados en diferentes Centros de Salud Mental de Bizkaia, pertenecientes a Osakidetza. El periodo de reclutamiento de la primera cohorte fue durante el año 2003, mientras que el de la segunda cohorte fue durante el año 2007.

Materiales y recogida de datos

Los pacientes de la cohorte 1 cumplimentaron el cuestionario específico Health Related Quality of Life for Eating Disorders version 2 (HeRQoLEDv2) (Las Hayas y cols., 2006 y 2007), el cuestionario genérico 12-item Short Form Health Survey (SF-12) (Gandek y cols., 1998; Ware y cols., 1996), y la versión española del Eating Attitudes Test-26 (EAT-26) (Castro y cols., 1991; Garner y cols., 1982). Un año más tarde, los pacientes volvieron a cumplimentar los mismos cuestionarios. Además, se recogieron datos sociodemográficos y datos clínicos mediante la revisión de las historias clínicas, y cumplimentación de un cuestionario clínico por parte de los psiquiatras que colaboraban en el estudio, tanto en el momento basal como en el seguimiento al año. Esta primera cohorte fue empleada para confirmar la existencia de una estructura interna de segundo orden con dos medidas resumen, y para la reducción de dicho cuestionario HeRQoLEDv2.

La cohorte 2 fue empleada para la validación de la versión reducida, Health Related Quality of Life for Eating Disorders-Short Version (HeRQoLED-S). Así, los pacientes de la cohorte 2 cumplimentaron la versión reducida del cuestionario específico (Las Hayas y cols., 2010), el cuestionario genérico SF-12 (Gandek y cols., 1998; Ware y cols., 1996), y la versión española del EAT-26 (Castro y cols., 1991; Garner y cols., 1982). Un año más tarde, los pacientes volvieron a cumplimentar los mismos cuestionarios. Además, se recogieron datos sociodemográficos y datos clínicos mediante la revisión de las historias clínicas, y la cumplimentación por parte de los psiquiatras colaborando en el estudio, tanto en el momento basal como en el seguimiento al año.

3.2. Análisis estadístico

3.2.1. Fiabilidad

Para evaluar la fiabilidad se estudió la consistencia interna de los diferentes instrumentos, que hace referencia a la homogeneidad de los ítems que componen una escala (Badia y Alonso, 2007). Para ello se estimó el coeficiente Alpha de Cronbach (Cronbach, 1951) para cada una de las escalas. Se consideró un resultado aceptable para valores superiores a 0,70 (Nunnally y Bernstein, 1994).

Por otro lado, con el fin de complementar los resultados de homogeneidad de los ítems que componen cada escala, se evaluó la validez convergente y discriminante de los ítems (concepto diferente a la validez convergente y discriminante de las escalas hasta ahora descrito) mediante la matriz de correlaciones de cada ítem con su escala y la matriz de correlaciones de cada ítem con el resto de las escalas. La validez convergente del ítem, se consideró adecuada si la correlación de un ítem con su escala tras ser eliminado dicho ítem de la escala (correcting for overlap) era superior o igual a 0,40. La validez discriminante del ítem se consideró adecuada si la correlación del ítem con su propia escala era significativamente mayor que con el resto de las escalas. La diferencia entre dos correlaciones se consideró significativa si dicha diferencia era de al menos dos errores estándar de la matriz de correlaciones, siendo el error estándar de la matriz de correlaciones, $1/\sqrt{n}$ (Fayers y Machin, 2000).

3.2.2. Validez de criterio

La validez de criterio únicamente se estudió para la versión reducida del cuestionario WOMAC en pacientes con artrosis de cadera. Se consideró como medida criterio o estándar oro la versión larga del cuestionario. Se evaluó la validez concurrente mediante las correlaciones entre las escalas del cuestionario

reducido y las escalas del cuestionario largo original mediante el coeficiente de correlación de Spearman.

Además se estudió el acuerdo entre las escalas reducidas y las escalas largas originales mediante el método de Bland-Altman (Bland y Altman, 1986), que nos permite detectar cualquier sesgo sistemático, evaluando el error aleatorio, y observando si las diferencias entre las puntuaciones de la versión reducida y la versión larga dependen del nivel de puntuación en las escalas.

3.2.3. Validez de constructo: validez convergente y discriminante

El estudio de la validez convergente y discriminante de las escalas se llevó a cabo mediante la correlación de las escalas del instrumento con las escalas de otros instrumentos. En cada caso se establecieron hipótesis previas en relación a qué escalas del instrumento a validar deberían estar altamente correlacionadas con qué escalas de otros instrumentos (validez convergente), y qué escalas del instrumento a validar deberían correlacionar bajo con qué escalas de otros instrumentos (validez discriminante). Se utilizaron los coeficientes de correlación de Pearson o Spearman. Se consideró un resultado aceptable si el coeficiente de correlación era superior a 0,40 con la escala hipotetizada, y menor que la propia consistencia interna de la escala (validez convergente), y si el resto de coeficientes de correlación eran inferiores a la correlación con la escala hipotetizada (validez discriminante) (Fayers y Machin, 2000).

3.2.4. Validez de constructo: validez de grupos conocidos

Para el estudio de la validez de grupos conocidos se compararon las puntuaciones de las escalas entre grupos de pacientes en los que se anticipaba que hubiera diferencias. Para la comparación de las puntuaciones según variable cualitativa de dos categorías (p. ej. sexo), se utilizó la prueba t de Student o el test no paramétrico de Wilcoxon. Para la comparación de las puntuaciones según variable cualitativa de más de dos categorías (p. ej. nivel de gravedad: leve, moderado o severo), se

utilizó el Análisis de la Varianza junto con las comparaciones múltiples de Scheffe, o el test no paramétrico de Kruskal-Wallis.

Además, se utilizaron los tamaños del efecto (TE) para medir la magnitud de la diferencia entre grupos (Cohen, 1992). El TE para la diferencia entre grupos se estimó mediante la siguiente fórmula:

$$TE = \frac{\bar{X}_{Grupo\ 1} - \bar{X}_{Grupo\ 2}}{DE_{Grupo\ 1 + Grupo\ 2}}$$

siendo DE la desviación estándar. Para clasificar la magnitud del TE se utilizaron los umbrales establecidos por Cohen (1992):

- Si TE < 0,20: no significativa
- 0,20 ≤ TE < 0,50: diferencia pequeña
- 0,50 ≤ TE < 0,80: diferencia moderada
- TE ≥ 0,80: diferencia grande

3.2.5. Validez de constructo: validez de estructura

La validez de estructura trata de realizar un análisis interno del instrumento para comprobar que la estructura hipotética subyacente del instrumento es correcta y adecuada, y para ello existen métodos tanto de la teoría clásica del test (TCT) como de la teoría de la respuesta al ítem (TRI).

Modelos factoriales de la teoría clásica del test

Entre los métodos de la TCT, se encuentran los modelos factoriales. La fórmula básica de la TCT es la siguiente:

$$Y_i = \eta + \varepsilon_i$$

donde Y_i es la puntuación del ítem i, η es el verdadero score del constructo a medir y ε_i es el término de error del ítem *i*. Estos modelos proporcionan combinaciones lineales de los ítems con los factores latentes que han de ser estimados mediante sumas ponderadas que reflejen la importancia de cada ítem en el constructo. Por tanto, el análisis factorial juega un papel muy importante dentro del estudio de la

validez de estructura de un instrumento. Las técnicas que hemos empleado son el Análisis Factorial Exploratorio (AFE), el Análisis Factorial Confirmatorio (AFC), los Modelos de Ecuaciones Estructurales (MEE), así como el caso específico de estos modelos para datos categóricos. Aunque los tres casos, AFE, AFC, MEE, se basan en detectar y analizar patrones en la matriz de correlaciones entre los ítems (o covarianza), y sirven para evaluar la dimensionalidad (número de factores) que es necesaria para representar la variabilidad que hay en los datos, el AFC y MEE permiten además contrastar estos patrones para confirmar la validez de los constructos hipotetizados. A continuación introduciremos brevemente los criterios utilizados para la evaluación de cada uno de estos modelos.

Análisis Factorial Exploratorio

El AFE trata de examinar la matriz de correlaciones de los ítems que componen el instrumento para así identificar grupos de ítems de forma que la correlación sea alta entre los ítems de un mismo grupo y baja con las de otros grupos. En el AFE no hace falta conocer a priori la estructura del instrumento, es por eso que se denomina exploratorio. En la Figura 3.1 se muestra un diagrama de flujo de un hipotético modelo factorial exploratorio, con siete ítems o variables observadas que se engloban en tres factores latentes o constructos. En el AFE, el investigador especifica el número de factores latentes, en este caso serían tres factores latentes, pero debe asumir ciertas condiciones:

a) todas las variables observadas o ítems están influenciadas por todos los factores;
b) cada variable observada o ítem está afectado por un factor único o error aleatorio;
c) todos los factores latentes, o están correlacionados entre sí, o son independientes; y
d) todos los factores únicos o errores aleatorios están incorrelacionados entre sí.

En el modelo planteado en la Figura 3.1, se observa como todos los ítems se ven afectados por todos los factores, de forma que decidiremos dependiendo del peso

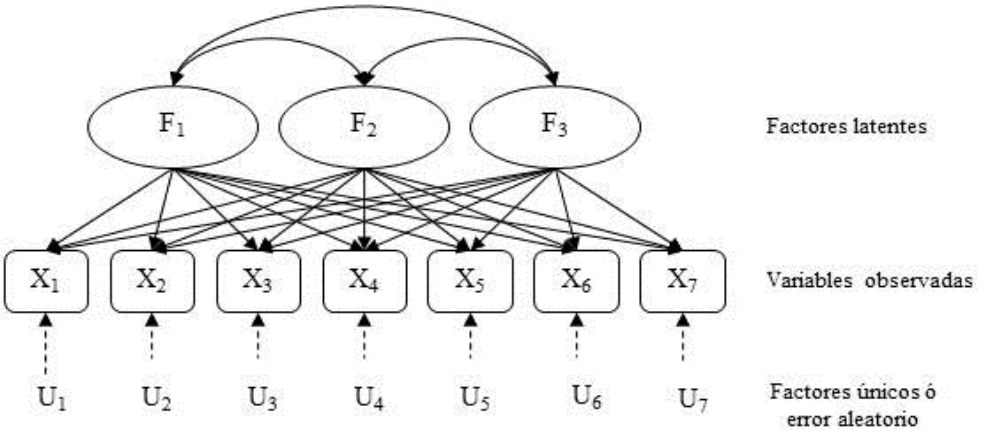


Figura 3.1. Diagrama de flujo de un modelo de análisis factorial exploratorio.

factorial de cada ítem en cada factor, como asignar los ítems a los factores. Además, se observa que hemos asumido que todos los factores latentes están correlacionados entre sí.

En nuestro caso, hemos utilizado el AFE para estudiar la estructura subyacente del cuestionario específico OP, que consta de ocho ítems que conforman un único factor, tal y como se muestra en el diagrama de flujo de la Figura 1.3 que se mostraba en el Capítulo 1 de la presente tesis. Los criterios que utilizamos para la evaluación del AFE fueron los siguientes: un ítem podía ser asignado al factor si su peso factorial era igual o superior a 0,40 y si su comunalidad era también superior a 0,40, siendo la comunalidad el grado en el que el ítem es representado a través del conjunto de factores considerados (Staquet y cols., 1998). En el Capítulo 4 se describe la metodología específica empleada con mayor detalle.

Análisis Factorial Confirmatorio

El AFC, a diferencia del AFE, trata de confirmar una estructura subyacente de un instrumento. Es decir, se establece a priori la estructura del instrumento (relación ítems y factores no observables) y a través de índices de bondad de ajuste se confirma si los datos se ajustan a dicha estructura. En la Figura 3.2 se muestra el

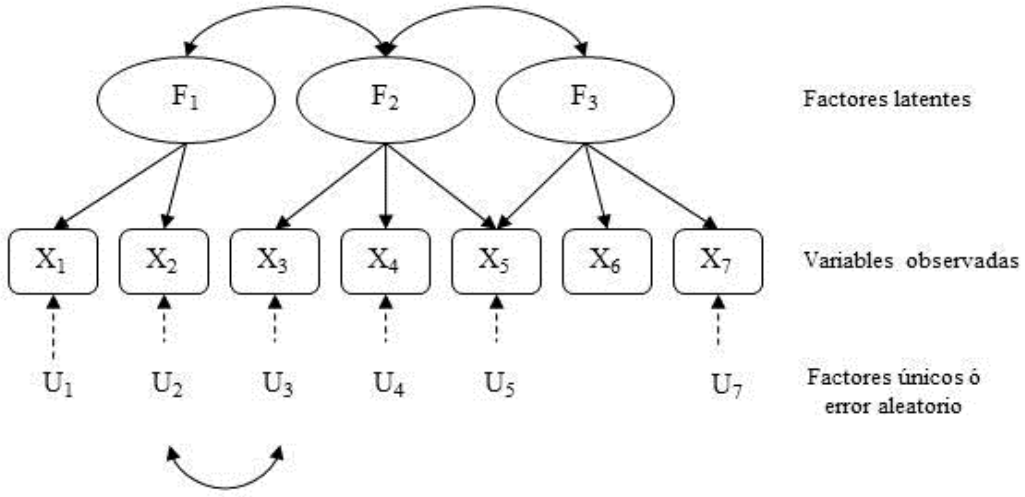


Figura 3.2. Diagrama de flujo de un modelo de análisis factorial confirmatorio.

diagrama de flujo de un hipotético modelo factorial confirmatorio, con siete variables observadas o ítems que se engloban en tres factores latentes o constructos. A diferencia del AFE, en el AFC el investigador debe considerar ciertos aspectos:

a) debe especificar que variables observadas o ítems están influenciadas por qué factores latentes;
b) debe especificar que variables observadas o ítems están afectadas por qué factores comunes o error aleatorio;
c) debe especificar la correlación o independencia entre los pares de factores latentes; y
d) debe especificar la correlación o independencia entre los pares de factores únicos o errores aleatorios.

En el modelo planteado en la Figura 3.2, se ha especificado que cada factor latente afecta solo a un conjunto de variables observadas o ítems. Además, se ha hipotetizado que el ítem X_6 no se ve afectado por ningún factor único (error aleatorio), además de que dos factores únicos están correlacionados entre si. Finalmente, se ha hipotetizado que los factores latentes F_1 y F_2, así como F_2 y F_3, están correlacionados entre sí, pero no los factores latentes F_1 y F_3. Así, una vez planteado el modelo, se estiman los parámetros, y mediante los diferentes test

estadísticos, así como medidas de bondad de ajuste se decide si el modelo planteado es adecuado o no.

En nuestro caso, se ha utilizado el AFC para estudiar la estructura subyacente tanto del cuestionario específico OP, cuya estructura se muestra en la Figura 1.3, como de la versión reducida propuesta del WOMAC, cuya estructura se muestra en la Figura 1.6. Aunque en el Capítulo 4 y Capítulo 5 se describe la metodología empleada con mayor detalle, a continuación resumiremos brevemente los criterios utilizados para la evaluación de los resultados del AFC.

Para la evaluación del modelo se emplearon diferentes índices de ajuste (Batista-Foguet y cols., 2004; Devins y cols., 2001; Hatcher, 1994; Mulaik y cols., 1989) para considerar si el modelo era adecuado:

- $\frac{\chi^2}{gl} < 2$, donde gl = grados de libertad.
- RMSEA (Root Mean Squared Error of approximation) < 0,08
- NNFI (Non-normed fit index) > 0,90
- CFI (Comparative fit index) > 0,90

Además, se estudiaron los pesos factoriales y se consideró un resultado aceptable si el peso factorial era ≥ 0,40, y significativamente distinto de 0. Por otro lado, se utilizaron los multiplicadores de Lagrange en caso de que el modelo necesitara modificaciones, ya que estos nos indican posibles relaciones que pueden ser añadidas para mejorar el ajuste del modelo. Se basa en identificar posibles parámetros o nuevas relaciones que podrían ser añadidas en el modelo para mejorar la bondad de ajuste del modelo. Estas relaciones pueden ser relaciones entre ítems y factores, o entre los propios factores. Consiste en estimar la reducción del estadístico χ^2 de bondad de ajuste al añadir una nueva relación, para ver cuánto puede mejorar el modelo. Reducir el valor del estadístico supondría mejorar el ajuste del modelo (Hatcher, 1994).

Modelo de Ecuaciones Estructurales

Los MEE son modelos factoriales que sirven para ajustar estructuras más complejas. Estos modelos también son modelos confirmatorios, siendo el AFC un caso particular de los MEE, con lo que hay que establecer a priori la estructura subyacente del instrumento. Una de las principales diferencias con el AFC es que estos modelos permiten tratar con variables causales, además de permitir ajustar estructuras más complejas como por ejemplo estructuras de segundo orden. En la Figura 3.3 se muestra un ejemplo de un MEE con estructura causal. Como vemos, se dispone de 17 ítems, que se conforman en cuatro factores de primer orden, "angustia psicológica", "dolor", "síntomas", y "náuseas y vómitos", además de un factor global de segundo orden denominado "calidad de vida". El modelo asume que una baja calidad de vida puede ser manifestada por angustia psicológica, dolor o síntomas. Y que por ejemplo, dicha angustia psicológica es una variable latente que tiende a resultar en ansiedad, tensión, depresión, irritabilidad, etc. Además, el modelo asume que los efectos adversos tales como náuseas y vómitos relacionados con el tratamiento, son indicadores causales y que son los que provocan un cambio en la calidad de vida.

En nuestro caso hemos utilizado los MEE para el cuestionario HeRQoLEDv2, con el fin de comprobar la existencia de una estructura de segundo orden con dos medidas resumen, tal y como se muestra en el diagrama de flujo de la Figura 1.7. Para la evaluación del modelo se emplearon los mismos índices de ajuste (Batista-Foguet y cols., 2004; Browne y Cudeck, 1992; Devins y cols., 2001; Hatcher, 1994; Mulaik y cols., 1989) empleados para el AFC, y los mismos criterios. Es decir, χ^2/gl < 2, donde *gl*=grados de libertad; RMSEA < 0,08; NNFI y CFI > 0,90; y pesos factoriales ≥ 0,40, y significativamente distintos de 0. También se utilizaron los multiplicadores de Lagrange para establecer nuevas relaciones que mejoraran el ajuste del modelo en caso de que el modelo necesitara alguna modificación (Hatcher, 1994). En el Capítulo 6 se describe la metodología específica empleada con mayor detalle.

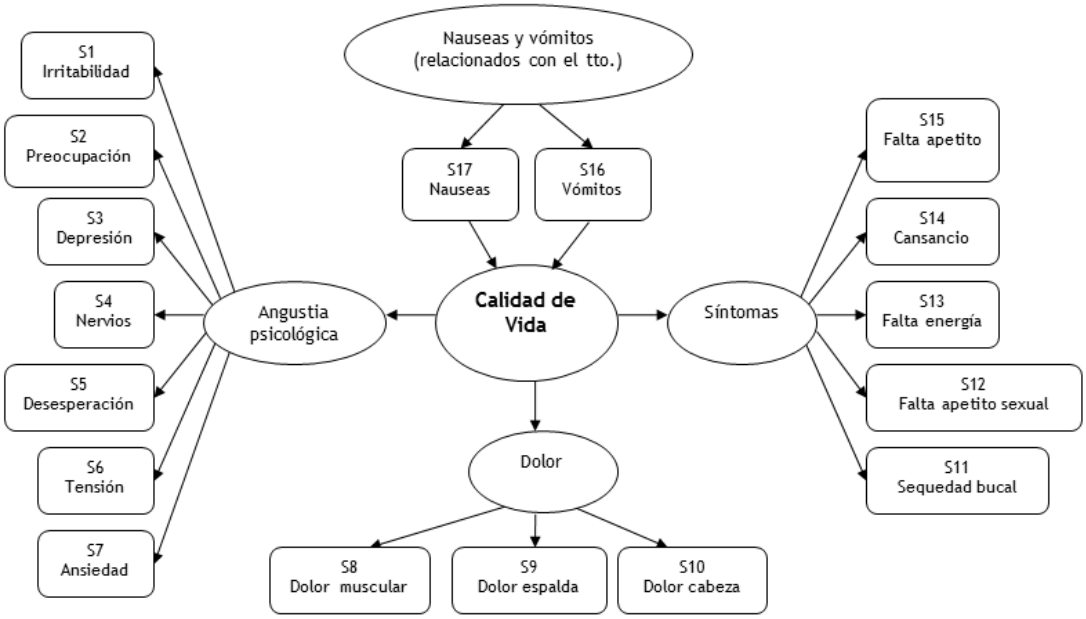


Figura 3.3. Diagrama de flujo de un modelo de ecuaciones estructurales (Fayers y Machin, 2007).

Modelo de Ecuaciones Estructurales para datos categóricos

Los MEE específicos para datos categóricos son un caso concreto de los MEE cuando los ítems son categóricos en vez de continuos. En estos modelos en vez de establecer la relación de los factores latentes sobre los ítems mediante modelos de regresión lineal se utilizan los modelos de regresión probit (Muthén y Muthén, 2010). Hemos utilizado los MEE para datos categóricos para la validación de la versión reducida del cuestionario HeRQoLEDv2. Una vez reducido el cuestionario, y llegado a la versión reducida HeRQoLED-S, utilizamos los MEE para datos categóricos con el fin de volver a comprobar si los ítems de esta versión reducida, además de conformarse en escalas de primer orden, también existe una estructura de segundo orden con dos medidas resumen tal y como se muestra en la Figura 1.8. Este análisis se realiza en una nueva muestra de pacientes con trastorno de la conducta de la alimentación.

Para la evaluación del MEE para datos categóricos se emplearon los siguientes índices de ajuste (Hu y Bentler, 1999; Mulaik y cols., 1989; Muthén y Muthén, 2010; Shevlin y Miles, 1998):

- $\frac{\chi^2}{gl} < 2$, donde *gl* = grados de libertad.
- RMSEA (Root Mean Squared Error of approximation) < 0,06
- TLI (Tucker-Lewis index) > 0,95
- CFI (Comparative fit index) > 0,95

Además, se estudiaron los pesos factoriales y se consideró un resultado aceptable si el peso factorial era ≥ 0,40, y significativamente distinto de 0. Por otro lado, se utilizaron los multiplicadores de Lagrange en caso de que el modelo necesitara modificaciones, ya que estos nos indican posibles relaciones que pueden ser añadidas para mejorar el ajuste del modelo (Hatcher, 1994). En el Capítulo 7 se describe la metodología específica empleada con mayor detalle.

Para la comparación de diferentes modelos, dado que estos no están anidados se utilizó el criterio AIC (Akaike Information Criterion), donde valores más bajos indican que el modelo ajusta mejor (Akaike, 1974).

Modelos de la teoría de la respuesta al ítem

Como ya se ha mencionado, dentro de los modelos de la TRI, existen modelos para ítems de respuesta dicotómica, siendo los más frecuentes los denominados modelos logísticos de 1, 2, o 3 parámetros, dependiendo del número de parámetros que se utilicen para su modelización. Los modelos logísticos de 1–parámetro se centran en el nivel de dificultad del ítem (Rasch, 1960). La dificultad del ítem se define como lo fácil o difícil que es para un respondedor aprobar ese ítem. Un ítem fácil se refiere a ítems que casi todo el mundo puede realizar dicho atributo (por ejemplo, levantar una mano). Por el contrario, un ítem difícil se refiere a un ítem donde muy poca gente es capaz de aprobarlo (por ejemplo, correr 10 km) (Huang y Speight, 2013). Cuanto mayor es el grado de dificultad mayor será la habilidad de la persona para responder positivamente a la pregunta. Los

modelos logísticos de 2–parámetros, además de la dificultad, también tienen en cuenta la discriminación del ítem. Ítems con mayor discriminación tendrán mayor capacidad de diferenciar niveles de habilidad de los individuos. Por tanto, los modelos de 2–parámetros son una generalización de los modelos de 1–parámetro que permiten diferente discriminación de los ítems (Hambleton y cols., 1991). Por último, los modelos logísticos de 3–parámetros, además de la dificultad y la discriminación del ítem, también consideran un tercer parámetro de incertidumbre que representa la probabilidad de que un individuo con baja habilidad responda de manera positiva a un ítem de gran dificultad. Este parámetro se introduce para controlar el llamado "adivinamiento" (Hambleton y cols., 1991). Cuando pasamos de un modelo logístico a otro es porque algún parámetro se fija a un valor concreto. Por ejemplo en el modelo de 2–parámetros se fija el parámetro de incertidumbre a 0. En el modelo de 1–parámetro, además, se fija el parámetro de discriminación de los ítems como constante, es decir, todos los ítems tiene la misma capacidad de discriminación (DeMars, 2010).

Por otro lado, existen extensiones de estos modelos de la TRI para ítems de respuesta politómica. Así, tanto el Partial Credit Model (Masters, 1982) como el Rating Scale Model (Andrich, 1978) son una extensión del modelo Rasch, donde el parámetro de discriminación es fijado como constante a 1. El Rating Scale Model es un caso particular del Partial Credit Model donde las distancias entre las diferentes respuestas de dificultad son iguales a lo largo de los ítems. Cada ítem tiene un nivel de dificultad o localización, pero se utiliza el mismo "step" de dificultades para indicar cada nivel (DeMars, 2010). Se asume que la transición de categoría a categoría es constante con lo que requiere que todos los ítems tengan el mismo número de opción de respuesta. Pueden ser de 3, de 4, de 5, o de 6 opciones de respuesta, pero siempre el mismo número. Para solventar este problema se creó el Partial Credit Model, que permite que la probabilidad de respuesta positiva de una categoría a otra varíe (Bond y Fox, 2007; de Ayala, 2009). Por otro lado, el Generalized Partial Credit Model (Muraki, 1992) es una generalización del Partial Credit Model en el que se permite que el parámetro de discriminación varíe a lo largo de los ítems. Se modeliza la probabilidad de respuesta a una categoría con

respecto a la categoría adyacente (de Ayala, 2009; DeMars, 2010). Otro método alternativo es el Graded Response Model (Samejima, 1969), en el que se modeliza la probabilidad de responder positivamente a una categoría de respuesta o una mayor. Por ejemplo, teniendo un ítem con tres opciones de respuesta, en los Partial Credit Models se modeliza la probabilidad de respuesta a la categoría 0 vs. la categoría 1, o la probabilidad de respuesta a la categoría 1 vs. la categoría 2. En los Graded Response Model de Samejima lo que se modeliza es la probabilidad de respuesta a 1 o mayor respecto 0. Es decir, la probabilidad de por debajo de una categoría con respecto a esa categoría o más alta. Son comparaciones acumulativas, y cada una de esas probabilidades acumuladas se modelizan con modelos dicotómicos. Es decir, los Graded Response Models especifican la probabilidad de respuesta a la categoría *k* o mayor vs. responder en categorías más bajas que *k* (de Ayala, 2009).

Entre los modelos de la TRI, teniendo en cuenta que los cuestionarios seleccionados están compuestos por ítems de respuesta politómica, hemos utilizado los métodos Rating Scale Model y Graded Response Model. A continuación introduciremos brevemente los criterios utilizados para la evaluación de cada uno de estos modelos.

Rating Scale Model

El Rating Scale Model es un modelo de 1–parámetro, que por tanto, trata de modelar la probabilidad de respuesta a un ítem en función de la dificultad que presenta el ítem y de la habilidad del individuo, asumiendo que la capacidad de discriminación de cada ítem es la misma, y que por lo tanto todos los ítems contribuyen de la misma forma al constructo que se está midiendo. En nuestro caso, se ha utilizado el Rating Scale Model para estudiar la estructura subyacente del cuestionario WOMAC reducido (Figura 1.6), con el fin de comprobar la unidimensionalidad de cada una de las sub-escalas de la versión reducida propuesta, dolor y función. Por tanto, se aplicó este análisis para cada una de las sub-escalas del cuestionario por separado. Además, se ha utilizado este mismo

modelo para reducir el cuestionario HeRQoLEDv2, eliminando ítems de nivel de dificultad similar o ítems redundantes. Es decir, se aplicó el Rating Scale Model de manera separada a cada uno de los dos factores de segundo orden del cuestionario, "Adaptabilidad social" y "Salud mental y rendimiento" (Figura 1.7).

Aunque en los Capítulos 5 y 6 se describe la metodología empleada con mayor detalle, a continuación describiremos brevemente los criterios utilizados para la evaluación de estos modelos:

a) La unidimensionalidad se comprobó mediante los estadísticos infit (mean square information-weighted statistics) y outfit (outlier sensitive statistic), considerando un resultado adecuado para valores comprendidos entre 0,7 y 1,3 (Tesio, 2003). Índices <0,7 son indicadores de ítems redundantes, e índices >1,3 son indicadores de ítems que no contribuyen al constructo. También se comprobó la unidimensionalidad mediante el análisis de componentes principales de los residuales del modelo (Davis y cols., 2003; Rothenfluh y cols., 2008; Smith y cols., 2006). Se consideró que no se cumplía la unidimensionalidad si además del primer factor, otros factores tenían valores propios mayores de 3 (Linacre, 2009).
b) Los índices de separación: el índice de separación de sujetos mide la capacidad del instrumento para distinguir o discriminar diferentes niveles de gravedad de los individuos. El índice de separación de los ítems mide la capacidad del instrumento de discriminar o distinguir diferentes niveles de dificultad (Rasch, 1960). Un índice de separación >2 es equivalente a una fiabilidad ≥0,80, y se considera aceptable (Duncan y cols., 2003).
c) Mapa de ítems-personas: es la representación de forma simultánea en la misma escala continua de los sujetos y los ítems ordenados según su nivel de habilidad y nivel de dificultad. Esta representación permite fácilmente visualizar la relación entre las habilidades de los sujetos y las dificultades de los ítems, para así poder comprobar si las dificultades de los ítems cubren todo el abanico de habilidad de los sujetos. Permite localizar áreas de habilidad no ajustadas por los ítems. Permite localizar clústeres de ítems

con similar nivel de dificultad. Permite ver si los ítems están ordenados de forma lógica según su contenido.

d) La independencia local: la independencia local indica que para sujetos con el mismo nivel de habilidad, las respuestas a un ítem son independientes de las respuestas a otro ítem. Es decir, para sujetos con el mismo nivel de habilidad, aunque se espere observar correlación entre cualquier par de ítems, esta debe surgir exclusivamente debido a que ambos ítems reflejan el mismo constructo. El supuesto de independencia local se comprobó mediante la correlación entre los residuos de todos los pares de ítems. Si esta correlación es <0,50 se asume independencia local (Davidson y cols., 2004).

e) La presencia de funcionamiento diferencial del ítem (DIF) se produce cuando un ítem presenta diferentes niveles de dificultad para diferentes grupos de individuos (Cook y cols., 2007) aun teniendo la misma habilidad respecto al rasgo que se pretende medir. Esta se comprobó comparando los diferentes niveles de dificultad de cada ítem según grupos definidos por algunas características sociodemográficas y clínicas. Se consideró que había DIF si el estadístico t de Welch resultó significativo, y si la diferencia entre los niveles de dificultad eran de al menos 0,50 logits (Linacre, 2009).

f) El funcionamiento de las categorías de respuesta de los ítems: se comprobó observando que los niveles de dificultad estimados para cada categoría de respuesta se presentaban de manera ordenada (Linacre, 2009; Reeve y cols., 2007).

Graded Response Model

El Graded Response Model es uno de los modelos TRI de 2–parámetros generalizado a datos politómicos ordenados (Samejima, 1969). A diferencia del modelo anterior, este incluye tanto el parámetro de dificultad como el de discriminación de los ítems (DeMars, 2010; Hambleton y cols., 1991), modelizando la probabilidad de responder positivamente a una categoría de respuesta o una mayor. Es decir, en el Graded Response Model se especifica la probabilidad de

respuesta a la categoría *k* o mayor vs. responder en categorías más bajas que *k* (de Ayala, 2009).

En el Graded Response Model, los parámetros del ítem son los siguientes: a) el nivel de dificultad (β), también denominado umbral, que nos indica el punto de la escala del rasgo latente en el cual una persona tiene una probabilidad de 0,50 de responder de forma positiva a una categoría del ítem; y b) la pendiente del ítem (α), que nos indica la habilidad del ítem para discriminar entre diferentes niveles de habilidad de individuos (Gomez, 2011). Por tanto, la respuesta del ítem viene conceptualizada en términos del parámetro de la pendiente (α), y una serie de *k*-1 umbrales o niveles de dificultad (β), donde *k* es el número de opciones de respuesta del ítem. Por ejemplo, en el caso de que los ítems tengan 4 opciones de respuesta, tendríamos 3 parámetros β para cada ítem siguiendo las siguientes respuestas dicotómicas: probabilidad de respuesta positiva a la primera categoría de respuesta frente al resto, probabilidad de respuesta positiva a las dos primeras opciones de respuesta frente al resto, y probabilidad de respuesta positiva a las tres primeras opciones de respuesta frente a la cuarta. De esta forma, cada ítem viene definido por un parámetro α (discriminación) y por *k*-1 parámetros β de dificultad. Y estos parámetros se utilizan para representar la Category Response Curve (CRC) para cada ítem, que representa la probabilidad de respuesta positiva a cada categoría de respuesta del ítem en función del rasgo latente (Gomez, 2008).

En nuestro caso, hemos aplicado el Graded Response Model a la versión reducida del cuestionario específico para personas con un trastorno de la conducta de la alimentación, el HeRQoLED-S cuya estructura se muestra en la Figura 1.8, para estudiar la estructura subyacente del cuestionario reducido, con el fin de comprobar la unidimensionalidad de cada una de las sub-escalas del cuestionario reducido, “Adaptabilidad social” y “Salud mental y rendimiento”. Por lo tanto, se aplicó este análisis a cada una de las sub-escalas del cuestionario por separado.

Para cada Graded Response Model, se estimó el parámetro de la pendiente (α), los parámetros de dificultad o umbrales (β), así como los errores estándar. Una mayor pendiente indica mayor capacidad de discriminación por parte del ítem. El rango de niveles de dificultad o umbrales establecidos por los parámetros β indican la adecuación del ítem para representar niveles bajos, medios o altos del constructo (Morrissey y cols., 2010). Además, se representó la curva de información del ítem (Item Information Curve) y la curva de información del test (Test Information Curve), que identifican la posición en el continuo latente en el que un ítem o la escala en general, respectivamente, proporciona más información (Lewis y Lambert, 2006).

Aunque en el Capítulo 7 se detalla la metodología empleada, a continuación describiremos brevemente los criterios utilizados para evaluar el ajuste del modelo (Gomez, 2008 y 2011; Lewis y Lambert, 2006):

a) La unidimensionalidad se comprobó mediante el AFE, donde un ratio de al menos 3:1 de los valores propios del primer y segundo factor sin rotar se consideró adecuado.
b) Se examinaron los errores estándar de las estimaciones de los parámetros, donde valores <0,20 se consideraron aceptables.
c) Se estimaron los residuales del modelo mediante la comparación de la proporción observada y estimada de cada categoría de respuesta, donde diferencias no significativas indicaban un buen ajuste.
d) Para cada ítem, se estimó la razón de verosimilitud, para evaluar la bondad de ajuste entre frecuencias observadas y esperadas, donde resultados no significativos indicaban un buen ajuste.
e) Para cada ítem se representó la CRC, que representa la probabilidad de respuesta positiva a cada categoría de respuesta del ítem en función del rasgo latente. Se consideró adecuada si se observa un cambio notable aumentando los niveles del constructo según aumentan las categorías de respuesta (Gomez, 2008).

3.2.6. Sensibilidad al cambio

Para el estudio de la sensibilidad al cambio, en primer lugar se compararon las características basales de los pacientes que habían respondido al cuestionario en el momento de seguimiento con respecto a los que no habían respondido, con el fin de comprobar si los no respondedores tenían algún perfil concreto. Para la comparación de las variables cualitativas se utilizó la prueba chi-cuadrado o el test exacto de Fisher. Para la comparación de variables cuantitativas se utilizó la prueba t de Student o el test no paramétrico de Wilcoxon.

Entre los respondedores al seguimiento, se estimaron los efectos techo y suelo de las escalas para evaluar la capacidad discriminatoria de estas. Se compararon las puntuaciones basales y las de seguimiento mediante la prueba t-test pareada o el test no paramétrico de los rangos con signo de Wilcoxon.

Para medir la sensibilidad al cambio de las escalas se utilizaron diferentes TE. Concretamente se utilizó el TE de Cohen (Standardized effect siez, SES) y la media de la respuesta estandarizada (Standardized response mean, SRM) (Guillemin y cols., 1993). Estos índices se estimaron mediante las siguientes fórmulas:

- $SES = \frac{\bar{X}_{Cambio}}{DE_{Basal}}$
- $SRM = \frac{\bar{X}_{Cambio}}{DE_{Cambio}}$

donde DE representa la desviación estándar. Para clasificar la magnitud de los TE se utilizaron los umbrales establecidos por Cohen descritos anteriormente (Cohen, 1992).

En los estudio en los que tras el seguimiento algunos pacientes podían mejorar y otros empeorar, se clasificó previamente a los pacientes según mejorados o empeorados y se realizó el estudio de sensibilidad al cambio de forma separada para cada grupo de pacientes.

3.2.7. Programas estadísticos

Para los análisis estadísticos se utilizaron diferentes programas. Para los análisis generales se utilizó el SAS for Windows statistical software (SAS Institute, Inc., carey; NC, USA). Para el análisis de MEE específico para datos categóricos se empleó el Mplus software (Muthén y Muthén, 2010). Para la aplicación de los modelos basados en la TRI, se utilizó el programa Winsteps (Linacre, 2009) en el caso del Rating Scale Model, y Mplus software (Muthén y Muthén, 2010) para el Graded Response Model.

Capítulo 4

Validation of the Spanish Translation of the Questionnaire for the Obesity-Related Problems Scale

Este capítulo ha sido publicado como investigación original en la revista Obesity Surgery (Factor de Impacto: 2,934; 1er cuartil), con referencia: Amaia Bilbao, Javier Mar, Blanca Mar, Arantzazu Arrospide, Gabriel Martínez de Aragón, José M Quintana. Validation of the Spanish Translation of the Questionnaire for the Obesity-Related Problems Scale. Obesity Surgery 2009;19(10):1393-400.

4.1. Abstract

Background: The objective of the study was to translate and validate the questionnaire for the Obesity-related Problems scale (OP) in the Spanish language.

Methods: The translation–retranslation procedure and a pilot study were first conducted. Then, patients with morbid obesity who were on a waiting list to undergo bariatric surgery were selected. The construct validity of the OP scale was studied by means of exploratory and confirmatory factor analyses. Item convergent and divergent validity were examined. The internal consistency of the OP scale, floor and ceiling effects, concurrent validity, and known-group validation were also examined. The Short Form 36 and EuroQol-5D questionnaires were used to evaluate the concurrent validity.

Results: A total of 123 individuals took part in the field study. Factor analyses and the item convergent and divergent validity confirmed the homogeneity of the construct. The Cronbach's alpha coefficient was high (0.93), and floor and ceiling effects were small. Regarding the OP scale concurrent validity, the majority exceeded the correlation coefficient of 0.40, and all correlation coefficients were lower than the internal consistency of the OP scale. In relation to the known-group validation, significant differences were found in the OP mean scale according to body mass index (BMI) and age. Those patients with a higher BMI and younger patients reported more obesity-related problems.

Conclusion: These results confirm that this version of the OP scale has been translated and adapted correctly for the Spanish language and that it fulfils the psychometric properties required for these instruments.

4.2. Introduction

Different studies of obese patients have indicated that obesity negatively affects health-related quality of life (HRQoL) as measured with generic or specific instruments (Fontaine et al., 2001; Kushner and Foster, 2000; Larsson et al., 2002). In the context of a study designed to quantitatively assess the impact of bariatric surgery on HRQoL in patients with morbid obesity, we used the specific Obesity-related Problems scale (OP) questionnaire in conjunction with two generic tools [EuroQol-5D (EQ-5D) and Short Form 36 (SF-36)] (EuroQoL Group, 1990; Karlsson et al., 2003; Sullivan et al., 1993; Ware and Sherbourne, 1992).

The OP scale is a self-assessment module developed in the Swedish Obese Subjects Study (SOS) to measure the impact of obesity on human psychosocial functioning (Karlsson et al., 2003; Sullivan et al., 1993). Because of the importance of psychological and physical functioning in the HRQoL of obese people and because the ultimate aim of our study was to quantify the improvement in HRQoL produced by the surgery, we selected the OP scale to see if weight loss associated with surgery improves psychosocial functioning. The other two generic HRQoL instruments have been translated into Spanish and validated (Alonso et al., 1995; Badia et al., 1999), but when our study began, we found no validated version of the Spanish translation of the OP scale questionnaire in the literature.

In adapting questionnaires from one language to another, standardized methodology is used to ensure that the instrument does not lose its psychometric properties (Aaronson et al., 1992; Gandek and Ware, 1998). The validity and reliability of an instrument can vary as a result of translation not only because of the linguistics but also because of the specific social and cultural features of the country. Thus, an assessment of the Spanish translation's properties before the routine use of this version with patients is necessary. The objective of this study was to show the translation and validation process of the Spanish version of the OP scale questionnaire.

4.3. Material and methods

4.3.1. Translation-retranslation procedure

According to the standards established by the International Quality of Life Assessment Project Group (Aaronson et al., 1992), the questionnaire was translated from the original language (English) into Spanish independently by two people whose native language was Spanish and who possessed a proficient level of English. The translations were compared by the coordinator of the study, differences were resolved by a consensus of the two translators, and a provisional forward translation was generated. This version was the basis for the retrograde translation process conducted by two native English speakers who were highly fluent in Spanish. Each person conducted a retranslation independently. These versions were evaluated by the coordinator and the two translators, compared with the original version, and then sent to the authors of the original questionnaire for review.

4.3.2. Pilot study

The Spanish version of the questionnaire was tested on a small sample of 20 patients with morbid obesity, who were on waiting lists to undergo bariatric surgery. The aim of this pilot study was to verify that the text of the Spanish version was not confusing or difficult to understand. The administration of the questionnaire was followed by a structured interview in which the patients individually assessed the characteristics of the different OP items. Patients were instructed to answer if any question was difficult to understand, difficult to answer, or confusing or whether they would phrase it in a different way. The final text, accompanied by a report on the translation and retranslation process, was sent for approval to the authors of the original version of the questionnaire.

4.3.3. Field Study

Interviews were conducted with patients diagnosed with morbid obesity who were waiting to undergo bariatric surgery at the Donostia (San Sebastian) and Txagorritxu (Vitoria) hospitals, both of which belong to the Basque Health Service-Osakidetza in the Basque Country. All 129 patients included on both waiting lists from May 2005 to May 2006 were asked to participate. Of those, four declined to answer the questionnaire, and two died before the interview. Therefore, the final sample was composed of 123 individuals who completed the interview. The study was approved by the hospital ethics committee.

The questionnaires administered were the OP scale, the SF-36, and the EQ-5D. The OP instrument (Karlsson et al., 2003; Sullivan et al., 1993) has eight questions, with four possible Likert-type answers that range from "definitely bothered" to "definitely not bothered", with two intermediate states (“mostly bothered” and “not so bothered”). The scale is recoded so that the worse the state the higher the score. Finally, the scores of all the questions are added to produce the raw total score, which can vary between 8 and 32. This score is standardized on a scale from 0 to 100. Thus, a 0 indicates a perfect state, and 100 the worst possible state.

The SF-36 and the EQ-5D are generic questionnaires applied to many different groups of patients, as well as the general population (Mar et al., 2005a and 2005b). The SF-36 (Ware and Sherbourne, 1992) has 36 items and covers eight domains (physical functioning, limitation of role activities because of physical problems, bodily pain, general health, vitality, social functioning, limitation of role activities because of emotional problems, and mental health) and two summary scales (physical and mental scales). The scores for the SF-36 domains range from 0 to 100, with a higher score indicating a better health status. The SF-36 has been translated into Spanish and validated in Spanish populations, and the measurement properties have been published elsewhere (Alonso et al., 1995). We also estimated utility values based on the SF-36, which is referred to as the SF-6D (Brazier et al., 2002).

The EQ-5D questionnaire (EuroQoL Group, 1990), a general health assessment instrument, has five questions about the subjects' state of health that measure mobility, self-care, performance of usual activities, pain or discomfort, and anxiety or depression. Each dimension is rated on a three-level scale from 1 (no problem) to 3 (inability to perform or extreme problem). The respondents' health status is expressed as a profile of their scores on each of the five questions. This study utilised the combined scores from the five questions. The combined dimensions describe 243 theoretically possible states of health that can be converted into a weighted health index score in which the higher the score the better the HRQoL. The EQ-5D has been shown to be reliable and valid, and it has been translated and validated in the Spanish population (Badia et al., 1998 and 1999).

In addition, general demographic and clinical data were also collected by trained personnel from the patients' medical records.

The patients were previously informed about the characteristics of the study, its voluntary nature, and the need to give their informed consent to be included in it.

4.3.4. Statistical analysis

The statistical description of the clinical and sociodemographic variables was carried out by frequency tables, means, and standard deviations (SD).

The construct validity of the OP questionnaire was studied by means of an exploratory factor analysis to test the hypothesis that the eight items on the questionnaire made up a single factor. An item was considered to be in the factor if the factor loading and communality were ≥ 0.40 (Staquet et al., 1998). To complement our results, a confirmatory factor analysis was also performed. Different fit indexes were evaluated (Batista-Foguet et al., 2004; Devins et al., 2001; Hatcher, 1994; Mulaik et al., 1989): (a) chi squared divided by the degrees of freedom, the result of which had to be ≤ 2 to be acceptable; (b) the root mean squared error of approximation (RMSEA), where a value <0.08 was considered

acceptable; and (c) the non-normed fit index (NNFI) and comparative fit index (CFI), both of which had to be >0.90 to be satisfactory. Factor loadings were also examined, and those ≥0.40 were considered acceptable. Therefore, if the model surpassed these acceptability criteria, it was considered acceptable.

The reliability was assessed with Cronbach's alpha coefficient (Cronbach, 1951). A coefficient >0.70 was considered acceptable for group data, and a coefficient >0.90 was recommended for individual assessment (Nunnally and Bernstein, 1994).

Item convergent and discriminant validity was examined by means of a matrix of item-scale correlations and correlations for each item with the other scales (SF-36, SF-6D, and EQ-5D). Item convergent validity was satisfied if the item-own scale correlation correcting for overlap was ≥0.40. Item discriminant validity was satisfied if the item correlated significantly higher with the scale it represented than with other scales. The significance of a difference between two item-scale correlations was determined by the standard error of the correlation matrix $\left(1/\sqrt{n}\right)$. The recommended significance criterion of two standard errors was used (Fayers and Machin, 2000).

The percentage of subjects scoring at the lowest possible level of the scale (floor effect) and the highest possible level (ceiling effect) were also examined. Floor and ceiling effects should be minimal, and we used 15% as the critical value for those effects (Wyrwich et al., 1999).

Concurrent validity was assessed by analysing the relationship between the OP scale and the domains of the SF-36, SF-6D, and EQ-5D with the Spearman correlation coefficient. We established that correlations between the OP scale and the other measures must be higher than 0.40 and lower than the internal consistency of the OP scale, as measured by Cronbach's alpha (Fayers and Machin, 2000). We also hypothesized that the correlation between the OP scale and the

social functioning domain of the SF-36 would be higher than with the other domains.

Known-groups validation was examined by comparing the OP scale among the different groups according to different criteria: (a) body mass index (BMI), comparing patients with a BMI ≤45 with those with a BMI >45, (b) age, comparing patients less than 50 years old with those 50 years of age or older, and (c) gender. We hypothesized that those patients with a higher BMI, women, and younger patients would report more obesity-related problems. Therefore, the OP mean score was compared among the different subgroups by the non-parametric Wilcoxon test. Further, to determine the magnitude of group differences, the effect sizes (ES) were calculated (Cohen, 1992). The ES of a between-group difference was estimated by calculating the mean difference divided by the pooled standard deviation. Cohen's benchmarks were used to classify the magnitude of effect sizes (Cohen, 1992): below 0.20 was not significant; between 0.20 and 0.50 small; between 0.50 and 0.80 moderate; and above 0.80 large.

Finally, we examined the relationship between the OP scale and mental well-being or mood disorders. To measure overall mood, we employed the mental health domain of the SF-36 (Wadden and Phelan, 2002). We examined the correlation between the overall mood measures and the OP scale with the Spearman correlation coefficient. Further, to illustrate the relation between the OP scale and the overall mood measures, patients were grouped into five categories according to the OP scores in the following manner (Karlsson et al., 2003): (a) no or very mild limitations in psychosocial functioning (0≤OP<20); (b) mild impact (20≤OP<40); (c) moderate impact (40≤OP<60); (d) severe impact (60≤OP<80); and (e) extreme impact (80≤OP≤100). Then, the mean score of the overall mood measure was compared among these five subgroups by means of the non-parametric Kruskal-Wallis test.

Effects were considered significant at $p<0.05$. All statistical analyses were performed with SAS for Windows statistical software, version 8.0 (SAS QC, 1999).

4.4. Results

The pilot study of the Spanish translation of the OP score conducted with 20 morbidly obese patients provided evidence of the acceptability and comprehensibility of the OP scale's questionnaire.

The sociodemographic and clinical data of the sample populations in the field study are shown in Table 4.1. The mean age was 42.93 years old (SD=10.70), and 85.37% of the sample were women.

Table 4.1. Characteristics of the patients (n = 123)

Characteristics	n	%
Gender		
Female	105	85.37
Male	18	14.63
Age, mean (SD)	42.93	10.70
BMI, mean (SD)	49.27	7.57
BMI groups		
< 40	5	4.07
40-44.9	37	30.08
45-49.9	32	26.02
50-54.9	30	24.39
55-59.9	9	7.32
≥ 60	10	8.13
Comorbidities		
Hypertension	40	32.52
Cardiopathy	4	3.25
Dyslipemy	20	16.26
Type 2 diabetes mellitus	24	19.51
OSAS	43	34.96

Data are given as number and percentage unless otherwise stated.
SD: standard deviation; BMI: body mass index; OSAS: obstructive sleep apnea syndrome.

The exploratory factor analysis of the eight items in the questionnaire (Table 4.2) showed factor loadings from 0.67 to 0.92 and item communalities from 0.45 to 0.85. In all cases, the benchmark of 0.40 was exceeded. The percentage of variance explained by the factor was 69.23%. Regarding the results of the confirmatory

factor analysis (Table 4.2), fit indexes were excellent: (a) chi squared divided by the degrees of freedom was 1.57, less than the benchmark of 2; (b) the RMSEA was 0.079, less than 0.08; and (c) the NNFI and CFI were 0.974 and 0.984 respectively, exceeding the benchmark of 0.90. Factor loadings were all statistically significant (p<0.001) ranging from 0.60 to 0.94.

Table 4.2. Results of the exploratory and confirmatory factor analyses and item convergent and discriminant validity

Items	Item description	Exploratory factor analysis		Confirmatory factor analysis	Item-scale correlations*	Correlations with others**
		Factor loading	Commu nality	Factor loading	Range of *r*	*r*
Item 1	Private gathering in my own home	0.78	0.61	0.71	0.30-0.58	0.76
Item 2	Private gathering in a friend's home	0.86	0.75	0.83	0.31-0.50	0.82
Item 3	Going to a restaurant	0.92	0.85	0.94	0.33-0.54	0.89
Item 4	Going to community activities	0.91	0.83	0.94	0.33-0.50	0.87
Item 5	Holidays away from home	0.90	0.81	0.87	0.27-0.62	0.85
Item 6	Trying on and buying clothes	0.67	0.45	0.60	0.20-0.37	0.57
Item 7	Bathing in public places	0.80	0.64	0.73	0.30-0.56	0.75
Item 8	Intimate relations with partner	0.78	0.61	0.72	0.33-0.56	0.70

* Item total correlation with its own OP scale correcting for overlap.

** Correlation between OP items and SF-36, SF-6D and EQ-5D domains.

r: Spearman correlation coefficient

With regard to reliability, the Cronbach's alpha coefficient was 0.93, highly superior to the minimum requirement of 0.70 for group data and above the recommended 0.90 for individual assessment.

The item-total correlations correcting for overlap ranged from 0.57 to 0.89, exceeding the benchmark of 0.40 (Table 4.2), and they were higher than the

correlation of each item with the other domains. On the other hand, the OP items correlated significantly higher with their own scale than with all other scales in 97.92% of the cases. In the remaining 2.08%, the difference between the correlation with their own scale and with the other scales was at least greater than 1.5 standard errors.

The correlation coefficients between the OP scale and the domains of the SF-36 ranged from -0.40 to -0.62, except for the role emotional domain, which was -0.34, and all were statistically significant (Table 4.3). The correlation coefficients with the SF-6D and EQ-5D were also satisfactory (-0.49 and -0.55, respectively) and significant (p<0.001), and they exceeded the minimum requirement of 0.40. Furthermore, all coefficients were lower than the Cronbach's alpha of the OP scale, and as hypothesized, the highest correlation coefficient was found with the social functioning domain of the SF-36.

Table 4.3. Obesity-related Problems scale (OP) correlation with SF-36, SF-6D and EQ-5D domains

Domains	OP scale
SF-36	
Physical functioning	-0.56*
Role physical	-0.48*
Bodily pain	-0.40*
General health	-0.51*
Vitality	-0.53*
Social functioning	-0.62*
Role emotional	-0.34*
Mental health	-0.53*
Summary physical component	-0.49*
Summary mental component	-0.48*
SF-6D	-0.49*
EQ-5D	-0.55*

Data are given as Spearman correlation coefficients.
* p<0.001

The mean OP scale was 53.74 (SD=31.81). The percentage of patients scoring at the lowest possible level of the scale (floor effect) and at the highest possible level

(ceiling effect) was consistently low (4.92% in both cases). Nonetheless, the critical value of 15% was not surpassed, and thus, it can be conclusively stated that there was not a ceiling or floor effect (Table 4.4).

Table 4.4. Known-groups validity of Obesity-related Problems scale (OP): mean scores for groups classified according to BMI, age, and gender

OP scale	Total	BMI		Age		Gender	
		<45	≥45	<50	≥50	Female	Male
n	123	42	81	84	39	105	18
Mean	53.74	45.14	58.25	57.50	45.43	53.46	55.46
SD	31.81	30.68	31.65	31.91	30.38	32.35	29.14
% at floor	4.92	7.14	3.75	4.76	5.26	5.71	0
% at ceiling	4.92	4.76	5	7.14	0	5.71	0
p value*		0.035		0.047		0.842	
ES**		0.41		0.38		0.06	

SD: standard deviation; BMI: body mass index; ES: effect size; % at floor: the percentage of the study population who score at the lowest possible scale level; % at ceiling: the percentage of the study population who score at the highest possible scale level.

* The non-parametric Wilcoxon test for the comparison of the OP scale between subgroups.

** ES was estimated by calculating the mean difference divided by the pooled standard deviation.

The differences in the OP mean scale were statistically significant between the two severity groups according to the BMI, and in relation to age (Table 4.4). Patients with a BMI of 45 or above had significantly higher scores on the OP scale compared with those with a BMI below 45 (p<0.05), and the magnitude of the difference was small (ES=0.41). Similarly, patients 50 years of age or older had significantly lower scores on the OP scale compared with those younger than 50 years of age (p<0.05), and in this case, the magnitude of the difference was lower (ES=0.38). In contrast, we did not find statistical differences in the OP scale scores according to gender (p=0.8420).

Finally, the correlation coefficient between the OP scale and the overall mood measure was -0.53 and was statistically significant (p<0.001). A significant decrease of the overall mood with increasing OP severity levels was demonstrated with comparison of the mean score of the overall mood measure among the five

subgroups, defined according to the OP scale; that is, the more the obesity-related psychosocial problems, the worse the general mental health (Table 4.5).

Table 4.5. Overall mood as measured by the mental health domain of the SF-36 of patients grouped according to OP scale score

Overall mood (SF-36 mental health domain)	Scores on the OP scale					*p* value*
	No or very mild problems (0-19.9)	Mild problems (20-39.9)	Moderate problems (40-59.9)	Severe problems (60-79.9)	Extreme problems (80-100)	
n	24	22	19	22	35	
Mean	76.33	66	58.95	52.18	43.66	<0.0001
SD	21.13	23.09	20.91	19.65	20.63	

SD: standard deviation.

* The non-parametric Kruskal-Wallis test for the comparison of the mental health domain of the SF-36 among subgroups defined according to the scores of the OP scale.

4.5. Discussion

This study shows that the Spanish version of the OP scale questionnaire is a valid and reliable instrument to measure the HRQoL of obese patients. Karlsson et al. studied the properties of this questionnaire in its original version (Karlsson et al., 2003) and observed that its psychometric properties were strongly supported. In this sense, our results reproduced the same good performance in terms of validity and reliability. Validity refers to the capacity to measure the concept that the instrument is designed to measure (Hair et al. 1998), and the reliability indicates that consistent results are generated (Staquet et al. 1998). Our aim was to translate and validate the obesity-specific OP scale in the Spanish language. As Fairclough pointed out (Fairclough, 2002), it is preferable to select a previously validated instrument than to create a new one. Our approach was oriented to make a cultural adaptation of a high quality instrument instead of creating a new one, because this one was unavailable in the required language, Spanish. Although other obesity-specific HRQoL instruments have been developed (Butler et al., 1999; Kolotkin et al., 1995 and 2001; Le Pen et al., 1998; Mannucci et al., 1999; Mathias et al., 1997;

Niero et al., 2002; Oria and Moorehead, 1998), in our opinion the OP scale questionnaire is short, easy to apply, and well understood by patients, and it provides a way to measure psychosocial functioning. The assessment of this domain has special meaning in these patients because of the risk of social marginalisation (Karlsson et al., 2003). In this sense, a key element in the assessment of the surgery-induced benefits is the measurement of improvement in psychosocial functioning. Other studies (Karlsson et al., 2003) have shown that the patients' engagement in new social activities correlates well with weight reduction. The use of the OP scale is useful to systematically evaluate the impact of surgery in the improvement of the social integration of patients. Weight loss is only a surrogate end point that allows an understanding of the progress of the treatment, but the final end points in the treatment of obesity are improvement in mortality rates and quality of life.

In this Spanish version of the OP scale, factor analysis results confirm the unidimensionality of the OP scale, given that the majority of the total variance is explained by the factor (69%), as found by the authors of the original questionnaire (Karlsson et al., 2003). Furthermore, we were able to show that the factor weights and communalities were all above the recommended threshold requirement of 0.40 (Staquet et al., 1998) and that they were grouped in a single area. Unlike the exploratory factor analysis, which calculates factor loadings on the basis of the actual data, the confirmatory factor analysis compares the actual data to an *a priori* model of how the data should look. Thus, confirmatory factor analysis is hypothesis-driven. The different tests performed in the confirmatory factor analysis provide strong support for the proposed structure of the OP scale.

The analysis of the internal consistency enabled us to confirm the hypothesis that the items that made up the area of the OP scale measured the same concept as the Cronbach's alpha coefficient, greatly exceeding the threshold of 0.70 established as the necessary limit for group data and the threshold of 0.90 for individual assessment (Nunnally and Bernstein, 1994). This indicates that the reliability of

the Spanish version of the questionnaire is as high as was found for the original questionnaire by its authors (Karlsson et al., 2003).

In addition, the item-scale correlations correcting for overlap show that the items are properly included in the OP scale, since all were by a wide margin above the threshold of 0.40 established as the necessary limit. The correlation coefficients were similar to those obtained by the authors of the original questionnaire (Karlsson et al., 2003). Therefore, item-convergent validity is demonstrated solidly—that is, all items contributed substantially to the total scale score. Furthermore, as some authors have pointed out (Karlsson et al., 2003), to assure that the OP scale is successful in measuring one concept as opposed to another, items should not be too closely associated with other constructs representing different domains (item-discriminant validity). In our study, the item-discriminant validity of the Spanish version of the OP scale was examined in relation to all other HRQoL measures, such as the SF-36, SF-6D, and EQ-5D. Although our study does not meet the criteria, as found by the authors of the original questionnaire (Karlsson et al., 2003), that all OP scale items correlated significantly higher with their own scale than with all other scales, it does meet the criteria in 97.92% of the cases, which is considered a very high percentage. Further, in those few cases in which the criterion is not met, the item-scale correlations were higher than correlations with the other scales, and the difference was at least higher than 1.5 standard errors. Thus, this pattern of results suggests an acceptable item discriminant and convergent validity.

The concurrent validity of the OP scale was assessed by examining the relationship between the OP scale and factors of the SF-36, SF-6D, and EQ-5D. All correlations were satisfactory, since all correlation coefficients were significant. Patients with psychosocial problems according to the OP scale had lower HRQoL levels in the SF-36 domains. Nevertheless, the OP scale had a lower correlation coefficient with the role emotional domain of the SF-36 (r=-0.32), and a higher correlation coefficient with the social functioning domain (r=-0.61). This last finding has also been reported by other authors (Kaukua et al., 2002). In any case, all correlation

coefficients between the OP scale and the collateral measures were moderate and lower than the Cronbach's alpha coefficient of the OP scale. As the authors of the original questionnaire have pointed out (Karlsson et al., 2003), this finding suggests that the OP construct provides unique information on the HRQoL of obese subjects. Therefore, correlations between the OP scale and collateral measures provide solid evidence for the concurrent validity of the Spanish version of the OP scale.

The mean OP score in our sample was 53.74, similar to that reported by other authors (Karlsson et al., 2003; Kaukua et al., 2002). Data for the floor and ceiling effects of the OP scale show that our results are similar to those in the original version (Karlsson et al., 2003), and, even so, they were negligible and never above the 15% considered as the critical value (Wyrwich et al., 1999).

Regarding the known-groups validity, we have shown that the OP scale is able to detect significant differences between BMI groups. The mean OP scale was higher in those patients with a higher BMI; that is, the impact of obesity on HRQoL varies directly with the severity of obesity, as other authors have pointed out (Fontaine et al., 1996; Karlsson et al., 2003; Kolotkin et al., 2001). The Spanish version of the OP scale also demonstrated the ability to differentiate between age groups. The mean OP scale was higher in those patients younger than 50 years of age, suggesting that the impact of obesity on psychosocial functioning is higher for younger patients. On the other hand, although other authors have pointed out that women report more obesity-related problems than men (Karlsson et al., 2003; Kolotkin et al., 2001), we could not demonstrate this finding probably because of the low number of men in our sample (14.63%). As found in other studies (Karlsson et al., 2003; Kolotkin et al., 2001), the percentage of women with obesity problems in our study was also much higher than the percentage of men with such problems. This fact has not allowed us to simultaneously study the OP score by gender and BMI. Therefore, the results of the known-groups validation suggest that the Spanish version of the OP scale detects differences among populations according to their BMI or age.

To assess the relationships between psychosocial functioning and mental well-being, we employed the mental health domain of the SF-36 as the overall mood measure (Wadden and Phelan, 2002). The correlation coefficient was similar to that obtained by the authors of the original questionnaire (Karlsson et al., 2003). In conclusion, high OP scores are related to poor overall mood.

Although this study showed the successful adaptation and subsequent validation of an instrument for measuring obesity-related quality of life into the Spanish language, there are some limitations to this study, such as the number of patients in the sample. Keeping in mind that the recommended minimum sample size to perform the analyses required to assess the appropriateness of the instrument (Fayers and Machin, 2000) is 5 to10 times the number of items, the size of the entire sample in our study was sufficiently large. However, because of the small sample size in some subgroups of patients, we must be careful when interpreting the results obtained from group comparisons. Thus, future studies should concentrate on validating the questionnaire in a wider sample. In addition, because an instrument must be reliable, valid, and responsive to be useful (Karlsson et al., 2003), assessment of the responsiveness of this Spanish version of the OP questionnaire in longitudinal studies is also necessary.

In conclusion, the obesity-specific questionnaire used in this study has demonstrated its reliability and validity in several ways. Although more analysis would strengthen our conclusions, our results confirm that this version of the OP scale has been translated and adapted correctly for the Spanish language and that it at least partially fulfils the psychometric properties required for these instruments.

Capítulo 5

Validation of a proposed WOMAC short form for patients with hip osteoarthritis

Este capítulo ha sido publicado como investigación original en la revista Health and Quality of Life Outcomes (Factor de Impacto: 2,112; 2º cuartil), con referencia: Amaia Bilbao, José M Quintana, Antonio Escobar, Carlota Las Hayas, Miren Orive. Validation of a proposed WOMAC short form for patients with hip osteoarthritis. Health and Quality of Life Outcomes 2011, 9:75.

5.1. Abstract

Background: The aims of this study were to propose a Spanish Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) short form based on previously shortened versions and to study its validity, reliability, and responsiveness for patients with hip osteoarthritis undergoing total hip replacement (THR).

Methods: Prospective observational study of two independent cohorts (788 and 445 patients, respectively). Patients completed the WOMAC and the Short Form (SF)-36 questionnaires before THR and 6 months afterward. Patients received the questionnaires by mailing, and two reminder letters were sent to patients who had not replied the questionnaire. Based on two studies from the literature, we selected the two shortened domains, the pain domain composed of three items and the function domain composed of eight items. Thus, we proposed an 11-items WOMAC short form. A complete validation process was performed, including confirmatory factor analysis (CFA) and Rasch analysis, and a study of reliability, responsiveness, and agreement measured by the Bland-Altman approach.

Results: The mean age was about 69 years and about 49% were women. CFA analyses confirmed the two-factor model. The pain and function domains fit the Rasch model. Stability was supported with similar results in both cohorts. Cronbach's alpha coefficients were high, 0.74 and 0.88. The highest correlations in convergent validity were found with the bodily pain and physical function SF-36 domains. Significant differences were found according to different pain and function severity scales, supporting known-groups validity. Responsiveness parameters showed large changes (effect sizes, 2.11 and 2.29). Agreement between the WOMAC long and short forms was adequate.

Conclusions: Since short questionnaires result in improved patient compliance and response rates, it is very useful to have a shortened WOMAC version with the same good psychometric properties as the original version. The Spanish WOMAC short

form is valid, reliable, and responsive for patients undergoing THR, and most importantly, the first WOMAC short version proposed in Spanish. Because of its simplicity and ease of application, the short form is a good alternative to the original WOMAC questionnaire and it would further enhance its acceptability and usefulness in clinical research, clinical trials, and in routine practice within the orthopaedic community.

5.2. Background

The disease-specific questionnaire, Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), is the most widely used instrument to evaluate symptomatology and function in patients with hip or knee osteoarthritis (OA) (Anderson et al., 1996; Bellamy et al., 1988a and 1988b; Hawker et al., 1995 and 1998). The measure was developed to evaluate clinically important, patient-relevant changes in health status resulting from treatment interventions (Bellamy, 2000). The WOMAC, which is self-administered and covers three dimensions: pain (5 items), stiffness (2 items), and physical function (17 items), is reliable, valid, and sensitive to changes in the health status of patients with hip or knee OA (Bellamy et al., 1988b; Escobar et al., 2002 and 2007b; McConnell et al., 2001; Quintana et al., 2005).

A major uses of health measurement scales is to detect health status changes over time, and a priority may be efficiency, i.e., responses achieved using the shortest possible questionnaire (Moran et al., 2001; Tubach et al., 2005). A shorter version would further enhance its applicability in epidemiologic studies, clinical trials, and daily clinical practice (Coste et al., 1997), since short questionnaires result in improved patient compliance and response rates and are thought to improve the quality of the response (Kalantar and Talley, 1999; Yang et al., 2007). Traditionally, one of the major disadvantages of self-administered questionnaires has been the low response rate, which greatly affects the study validity (Dillman, 1975; Kalantar and Talley, 1999), but it has been shown that shorter version of the questionnaires would significantly increase the response rate (Kalantar and Talley, 1999). In

addition, several studies have reported that the WOMAC function scale is redundant and suggested that the scale should be shortened by omitting the repetitious items (Ryser et al., 1999; Sun et al., 1997). Therefore, it would be very useful to have a shortened WOMAC version in Spanish, which retains the same good psychometric properties of the original version.

The WOMAC questionnaire has been shortened recently (Davis et al., 2003; Rothenfluh et al., 2008; Tubach et al., 2005; Whitehouse et al., 2003). Some have been shortened using statistical approaches (Davis et al., 2003; Rothenfluh et al., 2008), and others by considering the perspective of patients and rheumatologists (Tubach et al., 2005; Whitehouse et al., 2003). The stiffness domain of the WOMAC is largely redundant and is often excluded from the questionnaire (Whitehouse et al., 2003). Therefore, some authors have centred their studies on shortening the function domain (Tubach et al., 2005; Whitehouse et al., 2003), while others have shortened the pain and function domains (Davis et al., 2003; Rothenfluh et al., 2008), but these shortened domains have not been validated as a whole shortened WOMAC version, checking the existence of two underlying domains. Since the shortened scale is essentially a component of the fully shortened version, the subjacent structure of the reduced version should be analyzed.

The goal of the current study was to propose a shortened Spanish WOMAC version based on previously shortened versions and to evaluate the validity, reliability, and responsiveness of this shortened questionnaire for patients with hip OA, combining classical and modern statistical techniques, such as Rasch analysis.

5.3. Methods

5.3.1. Study population

The current study included data from two prospective cohorts recruited independently from various public teaching hospitals. Consecutive patients who underwent total hip replacement (THR) between March 1999 and March 2000,

and between September 2003 and September 2004, were eligible for the study and included in cohort 1 and 2, respectively. In both cohorts, patients with main diagnosis different to hip or knee osteoarthritis (OA), or with a malignant pathology or other organic or psychiatric condition that prevented participation, or with failure to undergo surgical intervention were excluded. Each hospital's ethics review board approved the study.

5.3.2. Measurements

The data collection and methodology for both cohorts were the same. All patients on the waiting list for a THR were mailed to their home address a letter that described the study and requested voluntary participation. The WOMAC (Bellamy et al., 1988b), short Form (SF)-36 (Ware and Sherbourne, 1992) questionnaires, and additional questions regarding the level of pain and function, which we will refer to as the categorical scales, were included in the mailing. The structure of those variables has been described previously (Quintana et al., 2000), and they classified patients as having minor, moderate, and severe pain or function. Therefore, patients completed the questionnaires at home, and they returned them by mail. A reminder letter was sent to patients who had not replied after 15 days. The patients who still had not responded after another 15 days received the questionnaire again and were contacted by telephone to ask them about the reasons of their non response. Six months after the intervention, patients received the same questionnaires and the follow up for those not responding was as described previously. Sociodemographic and clinical data also were collected.

The SF-36 is a generic questionnaire on health-related quality of life (Ware and Shebourne, 1992) that has 36 items and covers eight domains (physical function, physical role, bodily pain, general health, vitality, social function, emotional role, and mental health) and two summary scales on physical and mental health. The scores for the SF-36 domains range from 0 to 100, with higher scores indicating better health status. The SF-36 has been translated into Spanish and validated in Spanish populations (Alonso et al., 1995).

The WOMAC is a health status instrument specific for patients with hip or knee OA (Bellamy et al., 1988b). It has a multidimensional scale comprising 24 items grouped into three dimensions: pain (5 items), stiffness (2 items), and physical function (17 items). We used the Likert version of the WOMAC with five response levels for each item, representing different degrees of intensity that were scored from 0 (none) to 4 (extreme). The WOMAC has been translated into Spanish and validated in Spain (Escobar et al., 2002 and 2007b; Quintana et al., 2005).

After a thorough review of the literature and existing shortened WOMAC versions, we derived the WOMAC short form (WOMAC-SF) from the original WOMAC version to evaluate pain and function in patients with hip OA. The WOMAC pain short form was selected from a previously shortened version using Rasch analysis (Davis et al., 2003), which included items 1, 2, and 4 of the long form. The function short form included items 1, 2, 3, 6, 7, 8, 9, and 15 of the long form, selected from a previous study based on patients' and experts' opinions (Tubach et al., 2005). Some psychometric properties of the function short form have been investigated previously (Baron et al., 2007). Therefore, the WOMAC-SF that we proposed has 11 items grouped into two dimensions: pain (3 items) and function (8 items). The final scores for the long and short WOMAC versions were determined by adding the aggregate scores for pain and function separately, and standardizing them to a range of values from 0 to 100, with 0 representing the best health status possible and 100 the worst.

5.3.3. Statistical analysis

The unit of the study was the patient. In cases in which a patient underwent two interventions during the recruitment period, we selected the first intervention performed.

To describe the samples, we used means and standard deviations (SDs), frequencies, and percentages. We compared sociodemographic and clinical data and WOMAC domains at baseline between the cohorts. Chi-square or Fisher's exact

tests were performed to compare categorical variables, and the t-test or the Wilcoxon nonparametric test was used to compare continuous variables.

Cohort 1 was used to study all the psychometric properties performed to validate the Spanish 11-item WOMAC-SF. With the aim of studying the stability of items performance across different samples to give more evidence of validity, analyses concerning the construct validity were replicated in cohort 2.

Construct validity

We studied the construct validity by means of confirmatory factor analysis (CFA) to investigate the hypothesis that the 11 items on the questionnaire addressed two factors, pain and function. Different fit indexes were evaluated (Batista-Foguet et al., 2004; Devins et al., 2001; Hatcher, 1994; Mulaik et al., 1989): the root mean square error of approximation (RMSEA), for which a value below 0.08 was considered acceptable; and the non-normed fit index (NNFI) and comparative fit index (CFI), both of which had to exceed 0.90 to be satisfactory. We also examined factor loadings, and those 0.40 or higher were considered acceptable. We performed the CFA in both cohorts to study the stability of the subjacent structure of the questionnaire.

We applied the Rasch method to the WOMAC pain and function short forms separately to ensure that the scales were unidimensional (Rasch, 1960; Ryser et al., 1999), a fundamental requirement of construct validity (Streiner and Norman, 1989). We assessed unidimensionality by means of infit and outfit statistics, with values between 0.7 and 1.3 indicating a good fit (Tesio, 2003), and through a principal components analysis (PCA) of the residuals extracted from the Rasch model (Davis et al., 2003; Rothenfluh et al., 2008). Unidimensionality was considered violated if, in addition to the first factors, other factors had eigenvalues exceeding 3 (Linacre, 2009). We evaluated the ability of the WOMAC-SF to define a distinct hierarchy of items along the measured variable by means of an item separation index (Rasch, 1960). A value of 2.0 or greater for this statistic is

comparable to reliability of 0.80 and is acceptable. To detect the presence of differential item functioning (DIF), which occurs when different groups within the sample respond in a different manner to an individual item (Cook et al., 2007), we compared the different levels of the trait by gender. A Welch *t* statistically significant at $p<0.05$, and a difference in difficulty of at least 0.5 logit was considered as noticeable DIF (Linacre, 2009). We performed Rasch analyses in both cohorts to study the stability of the item logits and item order across the different samples.

Reliability

We assessed reliability using Cronbach's alpha coefficient (Cronbach, 1951). A coefficient over 0.70 was considered acceptable (Nunnally and Bernstein, 1994).

Convergent and discriminant validity

We assessed convergent and discriminant validity by analysing the relationship between the WOMAC-SF domains and the SF-36 domains with the Spearman correlation coefficient. We established that correlations between the WOMAC-SF domains and the other measures must be lower than the internal consistency of the WOMAC-SF scales (Fayers and Machin, 2000). We also hypothesized that the correlation between the WOMAC short pain scale and the bodily pain domain of the SF-36 and between the WOMAC short function scale and the physical function SF-36 domain would be higher than with the other domains.

Known-groups validity

We examined known-groups validation by comparing the WOMAC pain and function short scales among the different groups according to pain and function categorical scales (Quintana et al., 2000). We hypothesized that the more severe the patient's pain or function level, the higher their WOMAC pain and function short scores would be. Analysis of variance using the Scheffe test for multiple

comparisons or the non-parametric Kruskal-Wallis test was performed for the analysis.

Responsiveness

We compared principal characteristics between patients who responded to the follow-up and those who did not. Means and SDs were calculated for the WOMAC-SF scales at baseline and 6 months after surgery. We used a paired t-test for the comparison before and after the intervention. Ceiling and floor effects at baseline and 6 months after surgery were examined to evaluate the discriminatory ability of the scales.

To measure the responsiveness of the WOMAC-SF, we used the standardized effect size (SES), defined as the mean change score divided by the SD of the baseline scores, and standardized response mean (SRM), defined as the mean change score divided by the SD of the change scores (Guillemin et al., 1993). Cohen's benchmarks were used to classify the magnitude of the effect sizes (Cohen, 1992).

Agreement between the long and the short WOMAC forms

We evaluated the correlations between the pain and function long and short scales at baseline, 6 months after intervention, and for changes in scores by Spearman's correlation coefficient. Agreement between the WOMAC long and short scales was examined by the Bland-Altman approach (Bland and Altman, 1986), which is useful for searching for any systematic bias, assessing random error, and revealing whether the difference between the scores depends on the level of the scores (Baron et al., 2007).

All statistical analyses were performed with SAS for Windows statistical software, version 9.1 (SAS QC, 1999), except the Rasch analysis for which we used Winsteps version 3.69.1.4 software (Linacre, 2009).

5.4. Results

During the recruitment period, we included 788 and 445 patients in the first and second cohorts respectively, who underwent a THR, fulfilled selection criteria, and accepted to participate. Of these, 590 (74.87%) and 339 (76.18%), respectively, completed the questionnaires 6 months after the intervention. No differences were observed between both cohorts, except for the function categorical scale and WOMAC scales, with poorer results in cohort 2 (Table 5.1).

Construct validity

The results of the CFA for the hypothesized model of two latent factors, pain and function, provided satisfactory fit indices in both cohorts (Table 5.2). The RMSEA values were less than 0.08, and CFI and NNFI values were all exceeding the benchmark of 0.90. All factor loadings were significant (p<0.001) (range, 0.53-0.84) and similar in both cohorts, which supported the stability of the subjacent structure of the short questionnaire across the different samples.

Regarding the results of the Rasch analyses for the WOMAC pain and function short scales (Table 5.3), items were separated by 0.10 or more logit unit in both cohorts, which supported the stability of items across the different samples. Items were equally ranked based on their level of difficulty (δ) in both cohorts. Unidimensionality was supported with infit and outfit statistics ranging between 0.7 and 1.3, except the item “pain on sitting or lying” relative to pain scale in the first cohort (infit=1.33, outfit=1.32) and the item “putting on socks” relative to function scale in cohort 2 (infit=1.32). Furthermore, the PCA of the residuals did not yield additional factors with eigenvalues exceeding 3, since the second eigenvalue was 1.2 for the pain scale and 1.4 for the function scale in both cohorts, implying that the unidimensionality was not violated. In both cohorts, the person and item separation indexes exceeded 2, indicating reliability over 0.80. The presence of DIF by gender was not detected, given that in no case, the difference in

Table 5.1. Sociodemographic, clinical, and WOMAC preintervention descriptive statistics of samples

Parameter	Cohort 1 (n=788)	Cohort 2 (n=445)	*p* value
Age, mean (SD)	69.14 (8.91)	68.42 (9.81)	0.2039
Gender, women	381 (48.35)	221 (49.66)	0.6579
Body mass index			0.2707
<25	146 (19.36)	99 (23.19)	
25-30	358 (47.48)	198 (46.37)	
≥30	250 (33.16)	130 (30.44)	
Surgical risk			0.5047
ASA I-III	773 (98.10)	434 (97.53)	
ASA IV	15 (1.90)	11 (2.47)	
Charlson comorbidity index			0.9341
0	463 (58.76)	266 (59.78)	
1	218 (27.66)	121 (27.19)	
>1	107 (13.58)	58 (13.03)	
Pain categorical scale			0.4593
Minor	32 (4.09)	12 (2.72)	
Moderate	171 (21.87)	96 (21.77)	
Severe	579 (74.04)	333 (75.51)	
Functional limitation categorical scale			0.0076
Minor	79 (10.04)	36 (8.13)	
Moderate	422 (53.62)	206 (46.50)	
Severe	286 (36.34)	201 (45.37)	
WOMAC preintervention domains, mean (SD)			
Pain	54.27 (18.63)	58.16 (19.47)	0.0006
Function	65.19 (16.61)	68.44 (16.85)	0.0011

Data are expressed as frequency (percentage) unless otherwise stated.
Percentages exclude patients with missing data.
The scores for the WOMAC domains range from 0 to 100, with higher scores indicating worse health status.
SD = Standard deviation; ASA = American Society of Anesthesiologists.

the level of severity according to gender was statistically significant neither it was higher than 0.5 logits.

Table 5.2. Results of factor loading and fit indexes of Confirmatory Factor Analysis of the WOMAC short questionnaire in both cohorts

Items*	Item description	Cohort 1 (n=788)		Cohort 2 (n=445)	
		Pain	Function	Pain	Function
Pain item 1	Walking on flat surface	0.75	-	0.77	-
Pain item 2	Up/down stairs	0.84	-	0.84	-
Pain item 4	Sitting or lying	0.53	-	0.59	-
Function item 1	Descending stairs	-	0.74	-	0.74
Function item 2	Ascending stairs	-	0.74	-	0.77
Function item 3	Rising from sitting	-	0.67	-	0.67
Function item 6	Walking on flat surface	-	0.69	-	0.72
Function item 7	Getting in/out of a car	-	0.67	-	0.71
Function item 8	Shopping	-	0.71	-	0.70
Function item 9	Putting on socks	-	0.55	-	0.53
Function item 15	Getting on/off toilet	-	0.66	-	0.67
χ^2 (df)		226.11 (40)		119.97 (40)	
RMSEA		0.0792		0.0690	
CFI		0.9539		0.9650	
NNFI		0.9366		0.9518	

*Items are referred to by the original name in the WOMAC long form.

df = degrees of freedom; RMSEA = root mean square error of approximation; CFI = comparative fit index; NNFI = non-normed fit index.

Correlation between the two latent factors (pain and function) is set to be different from 0, therefore both latent factors are specified to be intercorrelated. The estimation of the correlation coefficient was 0.89 in the first cohort and 0.82 in the second one.

Covariance was specified between the error items of the following three pair of items: "Pain walking on flat surface" and "functional limitation walking on flat surface", "pain up/down stairs" and "functional limitation ascending stairs", and "functional limitation getting in/out of a car" and "functional limitation putting on socks".

Reliability

Cronbach's alpha coefficient was 0.74 for the WOMAC pain short scale, and 0.88 for the function short scale, which was superior to the minimum value of 0.70

Convergent and discriminant validity

The correlation coefficients between the WOMAC pain and function short scales and the SF-36 domains were all lower than the Cronbach's alpha of the WOMAC-SF

Table 5.3. Severity levels, standard errors, and goodness of fit indices of the pain and function short scales with application of the Rasch model in both cohorts

Items*	Item description	Cohort 1 (n=788)					Cohort 2 (n=445)				
		δ (logit)	SE	Infit MNSQ	Outfit MNSQ	Rank based on logit	δ (logit)	SE	Infit MNSQ	Outfit MNSQ	Rank based on logit
Pain†											
Item 4	Sitting or lying	2.21	0.07	1.33	1.32	1	2.30	0.09	1.30	1.29	1
Item 1	Walking on flat surface	-0.15	0.07	0.76	0.75	2	-0.07	0.09	0.84	0.87	2
Item 2	Up/down stairs	-2.06	0.07	0.88	0.89	3	-2.23	0.09	0.79	0.79	3
Function‡											
Item 6	Walking on flat surface	1.42	0.05	0.88	0.89	1	1.34	0.07	0.87	0.87	1
Item 15	Getting on/off toilet	0.83	0.05	1.15	1.14	2	0.96	0.07	1.16	1.15	2
Item 1	Descending stairs	0.63	0.05	1.01	0.99	3	0.37	0.07	1.00	0.97	3
Item 8	Shopping	0.01	0.06	1.07	1.04	4	0.01	0.08	1.08	1.04	4
Item 3	Rising from sitting	-0.15	0.06	0.93	0.95	5	-0.01	0.08	0.96	1.00	5
Item 2	Ascending stairs	-0.25	0.06	0.86	0.85	6	-0.36	0.08	0.84	0.79	6
Item 7	Getting in/out of car	-0.96	0.06	0.84	0.81	7	-0.91	0.08	0.85	0.82	7
Item 9	Putting on socks	-1.53	0.06	1.30	1.17	8	-1.41	0.09	1.32	1.22	8

*Items are referred to by the original name in the WOMAC long form.

†Cohort 1: person separation index=1.55; item separation index=24.98; cohort 2: person separation index=1.56; item separation index=19.44.

‡Cohort 1: person separation index=2.25; item separation index=15.39; cohort 2: person separation index=2.21; item separation index=10.44.

δ =level of severity (higher values indicate higher severity); SE=standard error; MNSQ=mean square fit statistic.

scales (Table 5.4). As hypothesized, the highest correlation coefficient of the WOMAC pain and function short scales were found with the SF-36 bodily pain and physical functioning domains respectively (-0.48 and -0.54).

Table 5.4. Correlation between the WOMAC short scales and SF-36 domains, and known-groups validity of the WOMAC short scales in cohort 1 (n=788)

SF-36 domains	WOMAC short scales	
	Pain ρ coefficient	Function ρ coefficient
Physical functioning	-0.44	-0.54
Role physical	-0.34	-0.36
Bodily pain	-0.48	-0.50
General health	-0.19	-0.17
Vitality	-0.32	-0.33
Social functioning	-0.38	-0.38
Role emotional	-0.17	-0.13
Mental health	-0.29	-0.25
Summary physical component	-0.34	-0.41
Summary mental component	-0.28	-0.25
	Pain Mean (SD)	Function Mean (SD)
Pain categorical scale		
Minor (n=32)[a]	24.74 (12.96)[b,c]	46.23 (17.56)[b,c]
Moderate (n=171)[b]	43.68 (14.16)[a,c]	57.97 (16.93)[a,c]
Severe (n=579)[c]	61 (17.16)[a,b]	72.67 (14.99)[a,b]
p value	<0.0001	<0.0001
Functional limitation categorical scale		
Minor (n=79)[a]	43.38 (16.63)[b,c]	54.06 (17.48)[b,c]
Moderate (n=422)[b]	51.70 (17.53)[a,c]	65.38 (15.70)[a,c]
Severe (n=286)[c]	64.94 (17.56)[a,b]	76.68 (15.43)[a,b]
p value	<0.0001	<0.0001

ρ: Spearman correlation coefficient.

Data are expressed as the Spearman correlation coefficient when studying the correlation between the WOMAC short scales and the SF-36 domains, and as the mean (SD) when comparing the WOMAC short scales according to the pain and functional limitation short categorical scales.

The scores for the WOMAC domains range from 0 to 100, with higher scores indicating worse health status. The scores for the SF-36 domains range from 0 to 100, with higher scores indicating better health status.

[abc] Superscript letters indicated differences among the three subgroups by Scheffe's test for multiple comparisons at $p<0.05$.

Known-groups validity

The differences in the WOMAC pain and function short mean scales were significant among the three severity groups according to the pain and function categorical scales (Table 5.4). Patients with a higher level of severity had significantly (p<0.0001) higher scores on the WOMAC pain or function short scale.

Responsiveness

There were no significant differences among the participants who responded to the follow-up and those who did not. Both the WOMAC pain and function short scales showed minor floor and ceiling effects (<2%) before the intervention (Table 5.5). After the intervention, the WOMAC pain and function short scales increased 39.28 and 39.99 points, respectively, both of which were significant (p<0.0001). The SES and SRM responsiveness parameters were much higher than 0.80 in both pain and function short scales, indicating large changes (Table 5.5).

Agreement between the long and short WOMAC forms

The long and short WOMAC scales at baseline, 6 months after the intervention, and the change scores were highly correlated (pain, r=0.94, 0.97, and 0.94, respectively; and function, r=0.95, 0.98, and 0.96, respectively). Agreement between the WOMAC long and short scales evaluated by the Bland-Altman approach is shown in Figure 5.1 and 5.2. For both domains, more than 95% of the differences between the two scales can be expected to be within the limits of agreement, and the variability was random and uniform along the range of values.

5.5. Discussion

The results of the current prospective study with two independent and large cohorts of patients who underwent THR at different hospitals and who were followed to 6 months support the validity, reliability, and responsiveness of the

Table 5.5. Responsiveness parameters 6 months after intervention in the WOMAC short scales in cohort 1 (n=590)

Parameters	WOMAC short scales	
	Pain	Function
% at floor		
Preintervention	0.68	0.17
Postintervention	31.21	6.24
% at ceiling		
Preintervention	1.88	1.89
Postintervention	0.17	0.17
Mean (SD)		
Preintervention	55.69 (18.64)	67.88 (17.44)
Postintervention	16.36 (17.95)	27.74 (19.48)
Change	39.28 (23.14)	39.99 (23.14)
p value*	<0.0001	<0.0001
SES	2.11	2.29
SRM	1.70	1.73

*Paired *t*-test to compare the mean preintervention and postintervention scores.
% at floor = percentage of the study population at the lowest possible scale level; % at ceiling = percentage of the study population at the highest possible scale level; SD = Standard deviation; SES = Standardised effect size; SRM = Standardised response mean.
The scores for the WOMAC domains range from 0 to 100, with higher scores indicating worse health status.
Changes were calculated by subtracting postintervention scores from preintervention scores; a positive result indicates a gain.

new 11-item version of the WOMAC. To the best of our knowledge, this is the first study to validate a shortened WOMAC version as a whole tool, including both pain and function dimensions, and most importantly, the first valid, reliable, and responsive WOMAC short version proposed in Spanish.

The WOMAC questionnaire is widely used both in research studies in orthopedic or rheumatologic processes as in clinical practice (Anderson et al., 1996; Bellamy et al., 1988a and 1988b; Hawker et al., 1995 and 1998; McConnell et al., 2001). One of the major disadvantages of self-administered questionnaires has been the burden of its completion (Aaronson et al., 2002). In some epidemiological and clinical studies, patients usually have to complete several questionnaires implying a great burden. In clinical practice, where information is collected to evaluate

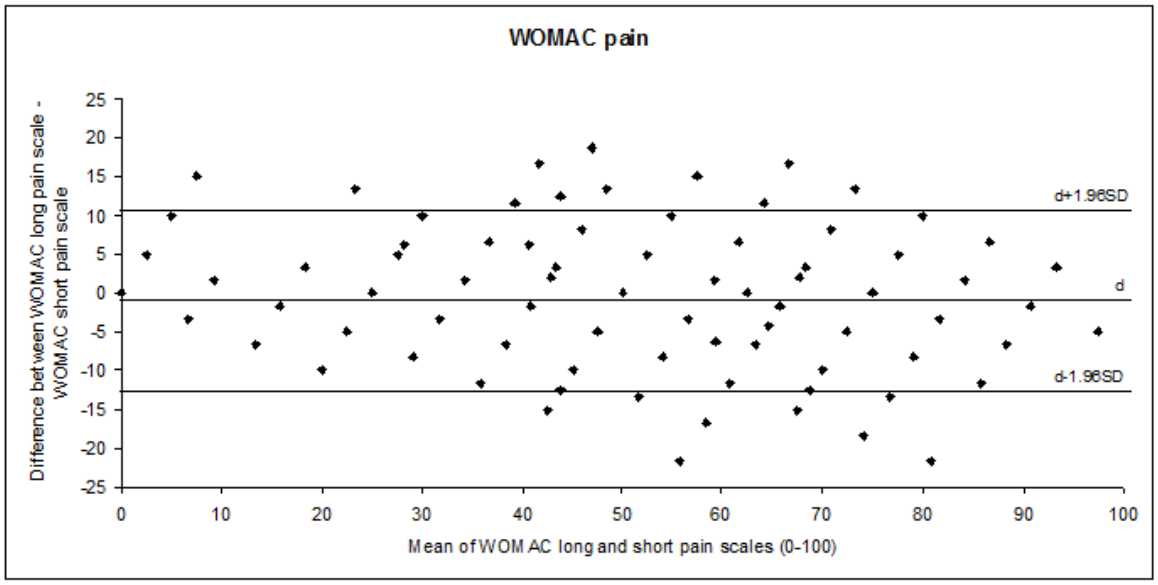


Figure 5.1. The Bland-Altman plot shows the difference in the WOMAC long and short pain scales plotted against the mean value of these two scales. The three horizontal lines indicate the mean individual differences d ± 1.96 SD (limits of agreement). The mean (SD) of the WOMAC long and short pain scales at baseline were 54.27 (18.63) and 55.70 (18.93), respectively. The mean (SD) of the difference between both scales was -1.47 (6.15). Limit of agreement: -13.52 to 10.58.

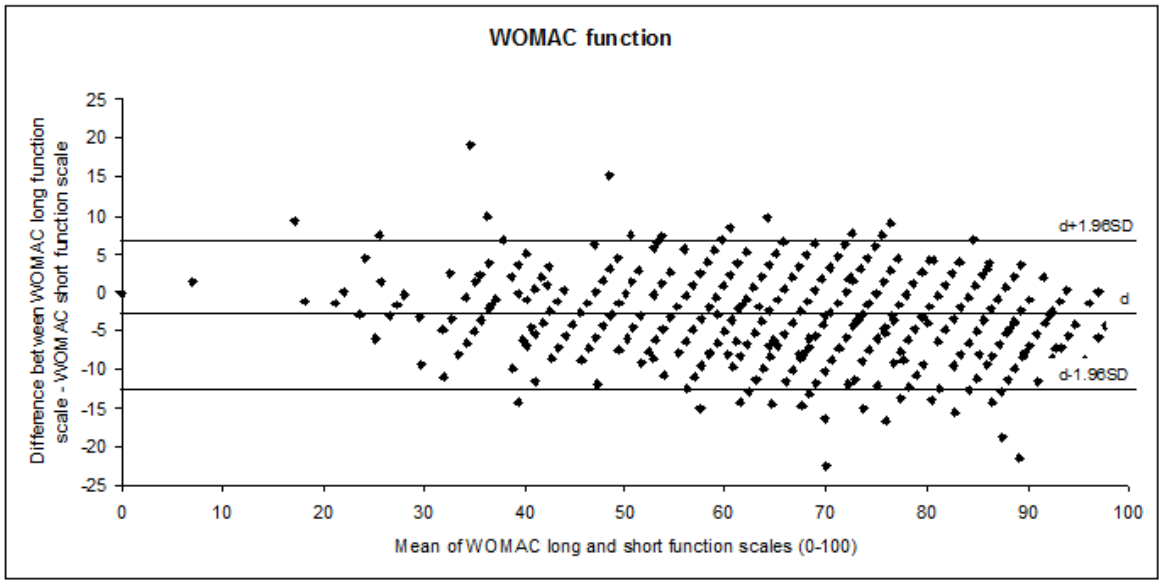


Figure 5.2. The Bland-Altman plot shows the difference in the WOMAC long and short function scales plotted against the mean value of these two scales. The three horizontal lines indicate the mean individual differences d ± 1.96 SD (limits of agreement). The mean (SD) of the WOMAC long and short function scales at baseline were 65.19 (16.61) and 68.36 (17.29), respectively. The mean (SD) of the difference between both scales was -3.15 (4.90). Limit of agreement: -12.75 to 6.45.

response to treatment, the goal is to involve as little effort as possible for both the patient and the physician. Therefore, if using a shortened version the same information is collected but with little burden, the instrument would be useful. In addition, another disadvantage of self-administered questionnaires has been the low response rate, which greatly affects the study validity (Dillman, 1975; Kalantar and Talley, 1999). Patients missing items has important implications for data collection, completion, and analysis. However, it has been shown that shorter version of the questionnaires would significantly increase the response rate (Kalantar and Talley, 1999), and the compliance increased when the respondent was asked to complete an appreciably smaller set of questions (Whitehouse et al., 2008). Therefore, a shorter version would further enhance its applicability in epidemiologic studies, and daily clinical practice (Coste et al., 1997). On the other hand, a consequence of the reduction of items is a loss in content validity, the comprehensiveness with which each domain is sampled, and investigators must be cognizant of this issue when they reduce the number of items (Moran et al., 2001). Because of a greater length of the questionnaire, it provides a detailed insight of different dimensions. However, this might also be a disadvantage, because of reduced patient compliance and incomplete response (Yang et al., 2007). Therefore, it would be very useful to have a shortened WOMAC version in Spanish, which retains the same good psychometric properties of its original version.

The aim of the current study was to propose a new short WOMAC form and validate it in Spanish. Fairclough (2002) commented that it is preferable to select a previously validated instrument than to create a new one. Considering this, and according to the different short versions of the WOMAC pain domain proposed by other investigators (Davis et al., 2003; Rothenfluh et al., 2008), we selected the shortened pain scale proposed by Davis et al. (2003). They shortened the WOMAC pain domain using Rasch analysis in a community sample of 773 patients with a hip or knee complaint. The authors concluded that the pain short scale fits the Rasch model and has interval-level scaling properties, and the stability of the model also was supported by a sample of 1,151 surgical patients. Rothenfluh et al. (2008) proposed a different three-item pain short version that had two items in

common with the version proposed by Davis et al. (2003), but the authors based it on a very small sample of patients with hip OA (n=57). Taking into account our objectives, the methodology used by Davis et al. (2003) for the reduction study, the larger sample size, and that both shortened pain domains had the same number of items, we decided that the pain short form proposed by Davis et al. (2003) was more adequate.

Regarding the WOMAC function short forms, other versions have been proposed by different authors (Davis et al., 2003; Rothenfluh et al., 2008; Tubach et al., 2005; Whitehouse et al., 2003). Davis et al. (2003), who based their new version on the Rasch model, also proposed a shortened version of the function scale. Nevertheless, they only excluded three items from the original version, and we did not consider short enough. Rothenfluh et al. (2008) also proposed a nine-item short version of the function scale based on the Rasch model but used a very small sample of patients with hip OA (n=57). Given that our target population is composed of patients with hip OA, we did not consider large enough the sample they used. Whitehouse et al. (2003) reduced the 17-item function scale to seven items by a clinically driven process based on the opinions of 36 orthopaedic and rheumatology personnel. The authors studied the validity, reliability, and responsiveness of the short scale in patients with hip or knee OA (Whitehouse et al., 2003), and the criterion validity and repeatability of this reduced function scale also was assessed in a sample of 100 patients, but only 30 had THR (Whitehouse et al., 2008). This short function scale also was validated in an independent cohort, but using a sample of patients with knee OA (Yang et al., 2007). Finally, Tubach et al. (2005), reduced the function scale from 17 items to eight, based on the opinions of 1,362 patients with hip or knee OA and 399 rheumatologists. This short function scale was validated in an independent sample of patients with hip or knee OA, and it was found to be responsive, reproducible, and valid (Baron et al., 2007). Although Whitehouse et al. (2003) and Tubach et al. (2005) used similar methods for shortening the scales, the latter considered more expert opinions, added patient opinions, and the scale was validated by also considering patients with hip

OA. Therefore, we selected the function short scale proposed by Tubach et al. (2005).

The validation studies of the various shortened WOMAC versions (Baron et al., 2007; Davis et al., 2003; Rothenfluh et al., 2008; Tubach et al., 2005; Whitehouse et al., 2003 and 2008; Yang et al., 2007) have consisted of studying the measurement properties of the corresponding shortened WOMAC pain or function scales individually. In our study, we validated our new 11-item WOMAC-SF as an entire tool, including both pain and function dimensions, and studying the construct validity of the short version to test the hypothesis that the 11 items in the questionnaire comprised two separate factors. Validation of the 11-item WOMAC-SF using CFA provides the questionnaire with greater construct validity. The CFA results confirmed the hypothesized internal structure of the two latent factors, given that all fit indices were satisfactory and all factor weights exceeded the recommended thresholds (Batista-Foguet et al., 2004; Devins et al., 2001; Hatcher, 1994; Mulaik et al., 1989). We also confirmed the internal structure of the 11-item WOMAC-SF by CFA performed in an independent cohort. A possible limitation could be the violation of the normal distribution of items when using the CFA. However, it has been argued that the maximum likelihood estimation procedure appear to be fairly robust against moderate violation of this assumption (Hatcher, 1994). In addition, some studies, based upon experience or computer simulations, have claimed that scales with as few as five points yield stable factors (Fayers and Machin, 2000). Therefore, taking into account that we use a 5-points Likert scale, a maximum likelihood estimator procedure, and that we have a large sample size, with practically equal results in both cohorts, we think that our CFA results are reliable and stable.

The Rasch method applied to the three-item pain short domain and the eight-item function short domain provided adjustment levels (infit and outfit) and unidimensionality sufficient to be considered adequate, providing major evidence of construct validity. Although two of the items, the item “pain on sitting or lying” relative to pain scale and the item “putting on socks” relative to function scale,

presented infit or outfit statistics slightly above the recommended threshold of 1.3, taking into account the satisfactory results obtained from the rest of analysis, such as PCA of the residuals, the functioning of the rating scale categories, the absence of DIF by gender in both items, and the item and person separation indexes, we do not consider that the slight difference in these infit or outfit indexes with respect to the recommended limit 1.3 is large enough to conclude that these two items are misfitting items. Regarding the three-item pain short form, the results were similar to those reported by Davis et al. (2003). Considering that the criteria were satisfactory, we concluded that the shortened WOMAC pain scale fit the Rasch model. Regarding the eight-item function short form, we obtained a scale that shows the fundamental properties of model fit and unidimensionality.

Analysis of the internal consistency allowed us to confirm the hypothesis that the items that comprised the pain short scale or those that comprised the function short scale measured the same concept as Cronbach's alpha coefficient exceeded the threshold of 0.70 (Nunnally and Bernstein, 1994). For the function short scale, the results were similar to or slightly higher than those reported by the original authors of the short form (Baron et al., 2007; Tubach et al., 2005). Further, the reliability of the 11-item WOMAC-SF, although it was as high as that for the original Spanish WOMAC questionnaire (0.82 for pain domain and 0.93 for function domain) because of the reduction of the number of items, it was slightly lower, indicating that it maintained excellent internal consistency (Escobar et al., 2002).

The convergent and discriminant validity of the WOMAC-SF was assessed by examining the relationship between the pain and function short scales and the factors of the SF-36. Validity was demonstrated by correlation coefficients lower than the internal consistency of the short forms and by confirming the hypothesis that the highest correlation coefficients were found between the WOMAC pain short form and the SF-36 bodily pain domain and between the WOMAC function short form and the physical function domain of the SF-36. Baron et al. (2007) also reported satisfactory convergent validity of the eight-item function WOMAC short form, but they used measures other than the SF-36. Whitehouse et al. (2003)

studied the convergent validity of their proposed seven-item function short form using the SF-36 physical function domain, and although the results were similar to those we obtained, in our case the correlation coefficient was slightly higher. Further, we obtained similar results to those of the original WOMAC questionnaire (Escobar et al., 2002), since they also found the highest correlation coefficient between the WOMAC pain and function long scales and the SF-36 bodily pain and physical functioning domains (-0.55 and -0.59, respectively). Otherwise, the WOMAC-SF maintained excellent known-group validity similar to that of the original WOMAC questionnaire (Escobar et al., 2002), since they also observed that the more severity level, the higher their WOMAC pain and function long scores were.

The 11-item WOMAC-SF showed good responsiveness 6 months after the intervention. Responsiveness parameters were substantially above the 0.80 threshold for designating large change (Cohen, 1992). Tubach et al. (2005) and Baron et al. (2007) also reported this finding for the function short form, although we found much higher responsiveness parameters, probably due to the follow-up period. We considered a follow-up of 6 months, whereas they considered 4 weeks. Whitehouse et al. (2003), who purposed a seven-item function WOMAC-SF, studied the responsiveness considering follow-up periods of 3 months and 1 year, and Yang et al. (2007), who validated the previous seven-item function WOMAC-SF in a different cohort, also studied the responsiveness considering follow-up periods of 3 and 6 months. Nevertheless, the responsiveness parameters of the seven-item function WOMAC-SF that they reported (Whitehouse et al., 2003; Yang et al., 2007) were much lower than our responsiveness parameters of the eight-item function short form that we proposed, indicating that the eight-item function short form is more responsive than the seven-item function short form proposed by Whitehouse et al. (2003). Further, the responsiveness results of the 11-item WOMAC-SF we obtained were similar to those of the original WOMAC questionnaire (Quintana et al., 2005), given that they also found minor floor and ceiling effects (<2%) before the intervention, and the SES and SRM responsiveness parameters were practically

equal (2.10 and 1.86 respectively for pain domain, and 2.34 and 1.80 respectively for function domain).

The strong correlation between the long and short WOMAC pain or function scales and the high agreement in scores examined by the Bland-Altman approach (Bland and Altman, 1986) support the hypothesis that the shortened scale captures pain and functional status as well as the original WOMAC version. Our results are similar to those found by Tubach et al. (2005) and Baron et al. (2007).

A possible limitation of the current study was the use of the data provided by the original WOMAC long form to validate the 11-item WOMAC-SF (Baron et al., 2007). This might constitute a framing bias and lead to overestimation of the similarity between the two forms (Baron et al., 2007; Whitehouse et al., 2003). Although this problem is inherent in many validation studies (Baron et al., 2007; Tubach et al., 2005), in the current study, whenever possible, we analyzed separate samples to compensate for this problem as much as possible. Nevertheless, the 11-item WOMAC-SF must be validated in a new independent sample of patients with hip OA and in different languages. Besides, the original WOMAC has been used in patients with hip or knee OA, consequently this 11-items short form could probably be applicable in both patients with hip or knee OA. However, we have based our study only on patients undergoing total hip replacement, and therefore, further validation studies in patients with different arthroplasties would be necessary to be completely sure about the applicability of this short WOMAC form.

In addition, an instrument must be reliable, valid, and responsive to be useful. Although we studied the reliability of the 11-item WOMAC-SF by means of the Cronbach alpha coefficient to measure the internal consistency, the reliability study should be complemented with a test-retest study. Regarding responsiveness, missing data are a key limitation of the prospective cohort design and a usual finding when conducting follow-up studies (Baron et al., 2007; Tubach et al., 2005; Whitehouse et al., 2003). In our case, there was a very good response rate before the intervention (about 80%) in both cohorts, and 6 months after it (about 75%).

These losses occurred despite our mailing up to two reminders and contacting nonresponders by telephone. However, no differences were observed in relevant variables when responders were compared with nonresponders. Therefore, although a bias may have been present in our responsiveness study due to missing data, it is likely to be minor and we believe the results are generalizable to the entire sample.

Conclusions

In conclusion, we proposed an 11-item WOMAC-SF, based on previous studies, for patients with hip OA undergoing THR. This complete validation process, which used two independent and large patient samples and combined classical and contemporary methods, such as Rasch analysis, showed that the 11-item Spanish WOMAC-SF is valid, reliable, and responsive for measuring pain and function in patients with hip OA undergoing THR, and most importantly, the first WOMAC short version proposed in Spanish. Its simplicity and easy of application will increase its acceptability and usefulness within the orthopaedic community, and, therefore, it may be of interest in routine practice given that the goal is to collect information involving as little effort as possible for both the patient and the physician. In clinical research, where patients usually have to complete several questionnaires implying a great burden, short questionnaires result in improved patient compliance and response rates, therefore this shorter version will further enhance its applicability. In conclusion, this short version is a good alternative to the original WOMAC questionnaire, since the 11-item WOMAC-SF retains properties of the original WOMAC version.

Capítulo 6

Use of Rasch methodology to develop a short version of the Health Related Quality of Life for Eating Disorders questionnaire: a prospective study

Este capítulo ha sido publicado como investigación original en la revista Health and Quality of Life Outcomes (Factor de Impacto: 1,860; 2º cuartil), con referencia: Carlota Las Hayas, José M Quintana, Jesús A Padierna, Amaia Bilbao, Pedro Muñoz. Use of Rasch methodology to develop a short version of the Health Related Quality of Life for Eating Disorders questionnaire: a prospective study. Health and Quality of Life Outcomes 2010, 8:29.

6.1. Abstract

Background: To confirm the internal structure of the Health Related Quality of Life for Eating Disorders version 2 questionnaire (HeRQoLEDv2) and create and validate a shortened version (HeRQoLED-S).

Methods: 324 patients with eating disorders were assessed at baseline and one year later (75.6% of whom responded). We performed a confirmatory factor analysis of the HeRQoLEDv2 using baseline data, and then a Rasch analysis to shorten the questionnaire. Data obtained at one year were used to confirm the structure of the HeRQoLED short form and evaluate its validity and reliability.

Results: Two latent second-order factors — social maladjustment and mental health and functionality — fit the data for the HeRQoLEDv2. Rasch analysis was computed separately for the two latent second-order factors and shortened the HeRQoLEDv2 to 20 items. Infit and outfit indices were acceptable, with the confirmatory factor analysis of the HeRQoLED short form giving a root mean square error of approximation of 0.07, a non-normed fit index and a comparative fit index exceeding 0.90. The validity was also supported by the correlation with the convergent measures: the social maladjustment factor correlated 0.82 with the dieting concern factor of the Eating Attitudes Test-26 and the mental health and functionality factor correlated -0.69 with the mental summary component of the Short Form-12. Cronbach alphas exceeded 0.89.

Conclusions: Two main factors, social maladjustment and mental health and functionality, explain the majority of HeRQoLEDv2 scores. The shortened version maintains good psychometric properties, though it must be validated in independent samples.

6.2. Background

Eating disorders (ED) affect millions of people worldwide. Since the earliest publications focusing on quality of life among individuals with an ED (de la Rie et al., 2005 and 2006; Doll et al., 2005; Keilen et al., 1994; Miller, 1996; Padierna et al., 2000 and 2002; Spitzer et al., 1995) it has been shown that they have a high degree of impairment in various areas of health-related quality of life (HRQoL). Most of these early studies used generic tools to assess the impact of an ED on physical, mental, and social factors (Hay and Mond, 2005). However, these generic tools did not include specific questions probing how the ED affected these factors which, in most cases, limited the interpretation of the results (Adair et al., 2007).

The first HRQoL instruments specific to individuals with an ED were published almost simultaneously in 2006 and 2007 (Abraham et al., 2006; Adair et al., 2007; Engel et al., 2006; Las Hayas et al., 2006 and 2007). We developed one of these, the Health Related Quality of Life for Eating Disorders version 2 (HeRQoLEDv2) questionnaire (Las Hayas et al., 2006 and 2007), a tool with good validity and reliability. One limitation of this 55- question instrument is that it requires a considerable amount of time to complete. We subsequently decided to develop a shorter version. Some techniques for shrinking the size of questionnaires arise from item response theory (IRT) (Beaton et al., 2005; Coste et al., 1997; Nijsten et al., 2006), with Rasch analysis being a useful approach. The rationale that makes Rasch models useful as a method to shorten the size of a questionnaire is that they can be employed to assess the unidimensionality of questionnaires, and remove items that disrupt this unidimensionality, identify degrees of trait severity and remove those items that overlap in severity level (Smith et al., 2006). In addition, it does not require large samples sizes for adequate parameter estimation (Linacre, 1994).

The objectives of the current study were to confirm a hypothesized internal structure of the HeRQoLEDv2, create a shortened version of this questionnaire (HeRQoLED-Short form), and then confirm the structure of the shortened version

and examine its validity and reliability. We hypothesized that the first-order factors of the HeRQoLEDv2 could represent two second-order latent traits: "social maladjustment" and "mental health and functionality." We tested this hypothesis in the present study.

6.3. Methods

6.3.1. Participants

Our detailed selection criteria have been described elsewhere (Las Hayas et al., 2006 and 2007). Briefly, the population consisted of ED patients being treated by four collaborating psychiatrists, experts in ED, working in three different mental health services in the province of Bizkaia, Spain. Diagnosis of an ED was performed by psychiatrists attending the patient if the patient met the diagnostic criteria for an ED established by the Diagnostic and Statistical Manual of Mental Disorders-IV (American Psychiatric Association, 1994).

Patients were excluded from the study if they had any serious multiorganic or psychotic disorder that could prevent adequate completion of the materials. To be included in the study, a patient had to participate in the investigation in an informed and voluntary way. The tenets of the Declaration of Helsinki were followed, and the study gained approval from the hospital's ethics committee.

Three questionnaires – the HeRQoLEDv2, the 12-item Short Form Health Survey (SF-12), and the Spanish version of the Eating Attitudes Test-26 (EAT-26) – were mailed to each patient's home address soon after recruitment, which we define as time 1 (T1). Those who did not respond in a timely fashion were sent reminders after 15 days and 30 days. The same questionnaires were mailed to patients one year later, which we define as time 2 (T2). As before, those who did not respond in a timely fashion were sent reminders after 15 days and 30 days.

Data from the T1 sample were used to perform confirmatory factor analysis (CFA) of the HeRQoLEDv2 followed by Rasch analysis. The T2 data were used to perform the CFA, validity, and reliability analyses of the shortened version.

6.3.2. Materials

Sociodemographic data were collected from each participant. In addition, each participant completed three selfadministered instruments related to HRQoL and ED.

The HeRQoLEDv2 (Las Hayas et al., 2006 and 2007) is comprised of 55 items and covering nine domains: symptoms, restrictive behaviors, body image, mental health, emotional role, physical role, personality traits, social relations, and binges. The scores in each domain are converted into a range from 0 to 100, with higher scores indicating a worse perception of HRQoL.

The SF-12 (Gandek et al., 1998; Ware et al., 1996) is a short generic survey of health status that can be summarized in two subscales: the physical component summary and the mental component summary. Values range from 0 to 100, with higher values indicating better health perception.

The Spanish version of the EAT-26 (Castro et al., 1991) was used as a measure of general eating disorder pathology. This test is composed of three factors – dieting concern, bulimia and food preoccupation, and oral control – and a total score. Its overall values range from 0 to 78, with higher scores indicating greater ED symptomatology.

6.3.3. Statistical analysis

Confirmatory factor analysis of the HeRQoLEDv2

The HeRQoLEDv2 had previously been submitted to an exploratory factor analysis to elucidate the way in which items relate to each other and with the hypothesized factors. Following this validity study (Las Hayas et al., 2006), we are now able to take a step further and hypothesize an internal structure of the HeRQoLEDv2 items and submit that structure to a confirmatory factor analysis. We excluded binges and symptoms domains from the model because binges domain was an independent domain and the symptoms domain is a list of symptoms rather than a proper measurement scale. A second-order CFA composed of a measurement model and a structural model was performed. We hypothesized a measurement model consisting of seven first-order factors: restrictive behaviors (6 items), body image (8 items), social relations (5 items), mental health (9 items), emotional role (4 items), physical role (4 items), and personality traits (4 items). These seven first order factors could be associated to two second-order latent traits: "social maladjustment" and "mental health and functionality". Based on both the content of the items from the following three first order factors "restrictive behaviours", "body image" and "social relations" and based on the literature, we believed that these three factors shared a common aspect: the impact of having an ED on the socio-cultural life. This impact is manifested in the way of feeding oneself, favouring the increase of restrictive behaviours and of feelings of body dissatisfaction (Stice, 1994). Also a recent study showed that families of individuals with ED perceived serious difficulties in their interpersonal relationship with the affected one (Hillege et al., 2006).

We also hypothesized that the mental health and functionality of individuals with an ED would affect their scores in the first-order domains of "physical role", "emotional role", "mental health", and personality traits". The mental health and functionality of ED individuals tend to be represented by a combination of high perfectionist traits, low self-efficacy feelings, stress due to feeling overweight and

depressive symptoms (Bardone-Cone et al., 2006 and 2008; Cowen et al., 1992). All of these traits and feelings are part of the content of the selected first order domains.

We further hypothesized that "social maladjustment" and "mental health and functionality" factors would be correlated given that an individual's mental state is likely to affect his or her social adjustment and vice versa.

Several different fit indices are applicable to these analyses (Browne and Cudeck, 1992; Hatcher, 1994). We used the chi-square test divided by degrees of freedom, the results of which had to be less than 2.0 to be acceptable (Hatcher, 1994); the root mean square error of approximation, where values of 0.08 or less are acceptable (Browne and Cudeck, 1992); and the non-normed fit index and comparative fit index, both of which had to be equal to or greater than 0.90 to be satisfactory (Hatcher, 1994).

Only items that showed factor loadings ≥ 0.40 in the corresponding factor were accepted (Hatcher, 1994). The Lagrange multiplier test, which identifies paths or covariances that should possibly be added to the model to improve the fit, was used when the model needed modification.

CFAs were performed with the CALIS procedure of the SAS program (version 8.0) (SAS QC, 1999).

Rasch analysis

The Rasch method was applied to the original version of the HeRQoLED as a means to develop the Health Related Quality of life for Eating Disorders - Short Form (HeRQoLED-S). The Rasch model presumes that a single trait drives item responses (Cook et al., 2007), so that a person's response to an item that measures a single trait is accounted for by his/her level (amount) on that trait, and not by other factors (Reeve et al., 2007). The Rasch model assumes that the probability of

a given patient responding affirmatively an item is a logistic function of the relative distance between the item location parameter (the difficulty of the item) and the respondent (the ability of the patient), and only a function of that difference (Prieto and Delgado, 2003). Items along the logit scale are ordered according to its difficulty level; the most difficult ones are at the top and the easiest ones, at the bottom (Wolfe and Kong, 1999). In our study, items which reflect the highest impact on HRQoL are placed at the top of the continuum and those which reflect the lowest impact are placed at the bottom. We used the polytomous Rasch rating scale model because our response scales are ordinal with six response options. A joint maximum-likelihood estimation process was used to estimate the parameters (Wright and Stone, 1979).

Prior to all further analyses, the functioning of rating scale categories was examined for each of the two domains of the HeRQoLED short form. The rating scale categorizations presented to respondents are intended to elicit from those respondents unambiguous, ordinal indications of the locations of those respondents along the latent trait of interest (Linacre, 2009). Therefore the probability of selecting an item response category indicative of better health status should increase as the underlying level of health of the respondent increases (Reeve et al., 2007). Linacre (2009) suggests the following criteria to assess adequate functioning of rating scale categories: (1) More than 10 observations per category (or the findings may be unstable, i.e., nonreplicable); (2) A smooth distribution of category frequencies. The frequency distribution is not jagged; (3) Clearly advancing average measures; (4) Average measures near their expected values; (5) Observational fit of the observations with their categories: Outfit mean squares near 1.0. Values much above 1.0 are much more problematic than values much below 1.0.

Because the condition of unidimensionality is a requirement for using Rasch analysis, we applied the Rasch analysis separately to both social maladjustment and mental health and functionality factors. Unidimensionality was assessed through a principal components analysis (PCA) of the residuals extracted from the

Rasch model (Smith et al., 2006). A violation of unidimensionality was considered if in addition to the first factor there were other factors with eigenvalues greater than 3 (Linacre, 2009). Apart of the PCA, unidimensionality was assessed through examination of fit statistics. We used two indices of fit, namely the mean square information-weighted statistic (infit) and the outlier-sensitive statistic (outfit). Values between 0.7 and 1.3 for both indices indicate a good fit (Tesio, 2003).

We evaluated how well the HeRQoLED - short version differentiates individuals in the measured domains on the basis of the person separation statistic (Duncan et al., 2003) and how well it differentiates items based on the item separation index, which indicates the ability to define a distinct hierarchy of items along the measured variable. A value ≥ 2.0 for this statistic is comparable to a reliability of 0.80 and is acceptable. Correlation of items with the total scale score served to evaluate whether the items correlated in a similar way with the construct being measured (Cole et al., 2004).

"Item bias" or "differential item functioning" (DIF) occurs when items exhibit different difficulties for different person groups. For a given level of a trait, the probability of endorsing a specified item response should be independent of group membership (Cook et al., 2007). For the DIF analysis, we examined whether diagnosis subtype (anorexia nervosa, bulimia nervosa, or eating disorder not otherwise specified) may exert influence on item calibrations in subsamples. DIF analyses were performed independently for the "Social maladjustment scale" and for the "Mental health and functionality scale". Welch t gives the DIF significance as a Welch's (Student's) t-statistic. The t-test is a two-sided test for the difference between two means (i.e., the estimates) based on the standard error of the means (i.e., the standard error of the estimates). The null hypothesis is that the two estimates are the same except for measurement error. To establish a noticeable DIF between subsamples, the difference in difficulty of the item between the two groups (DIF contrast) should be at least 0.5 logits. In addition, the Welch t should be statistically significant, $p < 0.05$ (Linacre, 2009).

Residual correlations between items within a scale were examined for local dependency. Correlations > 0.5 between item residuals can indicate that responses to one item may be determined by those to another (Davidson et al., 2004).

Rasch analyses were repeated until we obtained a version that met the criteria, which was named the Health Related Quality of Life for Eating Disorders-Short Form (HeRQoLED-S). Item content was examined for the misfitting items before removal from the scale. Two of the authors of the present study (JAP and CLH) are experts in the field of eating disorders. They jointly decided whether to retain or delete an item based on the clinical importance of the content. Winsteps version 3.37 was used for the Rasch analysis (Linacre, 1999).

Confirmatory factor analysis of the HeRQoLED-S

A CFA was applied to the shortened version. The hypothesized structural and measurement models were the same as those of the long version. The only difference was that fewer items were assigned to each first order factor. The same fit indices were also used to assess the goodness of fit.

Validity and reliability of the HeRQoLED-S

Based on content similarity between subscales of different questionnaires, we hypothesised the following correlations for the analysis of concurrent validity: The social maladjustment factor would correlate positively and moderately, by means of the Pearson correlation coefficient, with the dieting concern factor of the EAT-26. The mental health and functionality factor, in turn, was hypothesized to correlate negatively and moderately with the mental component summary of the SF-12. The Cronbach alpha index of reliability was calculated for each factor; values above 0.70 were acceptable (Cronbach, 1951).

6.4. Results

Participants

A total of 394 ED patients were approached for the study. Of them, 324 ED patients completed the first set of questionnaires (T1). All patients were receiving treatment for their ED at T1 but they differed in ED subtype, severity and time in treatment. We did not filter patients in these regards; therefore we expect that these patients represent the entire spectrum of ED severity. All were asked to complete the same tests again after one year. Of these, 245 patients (75.6%) responded. Most participants were women (96.3% at T1 and 95.1% at T2), with a mean age of 27 years, SD (8.76) at T1. From the baseline sample, 21% patients had been diagnosed with anorexia nervosa, 15% with bulimia nervosa, and 64% with eating disorders not otherwise specified.

Confirmatory Factor Analysis of the HeRQoLEDv2

For the CFA, only data from the 262 participants who completed the HeRQoLEDv2 at T1 with no answers missing were used. The hypothesized model described in the Introduction provided satisfactory fit indices after few adjustments. Following the Lagrange multiplier test, two pairs of errors, one belonging to the body image domain and the other to the social relations domain, were allowed to covary. Additionally, the Lagrange multiplier test suggested setting a new causal relationship between the personality traits item "Have you had lack of confidence in your own capabilities?" and the mental health domain item "Have you felt yourself worthless?". This new relation is meaningful given that lack of confidence in one's capabilities may lead an individual with an ED to feelings of worthlessness when facing problems. After these adjustments, the goodness of fit indices for the model were satisfactory (χ^2 (df = 729) = 1464.67, P < 0.0001; χ^2/df = 2.01; RMSEA = 0.06; NNFI = 0.90 and CFI = 0.90). Figure 6.1 shows the path diagram of the model with the estimated parameter values included.

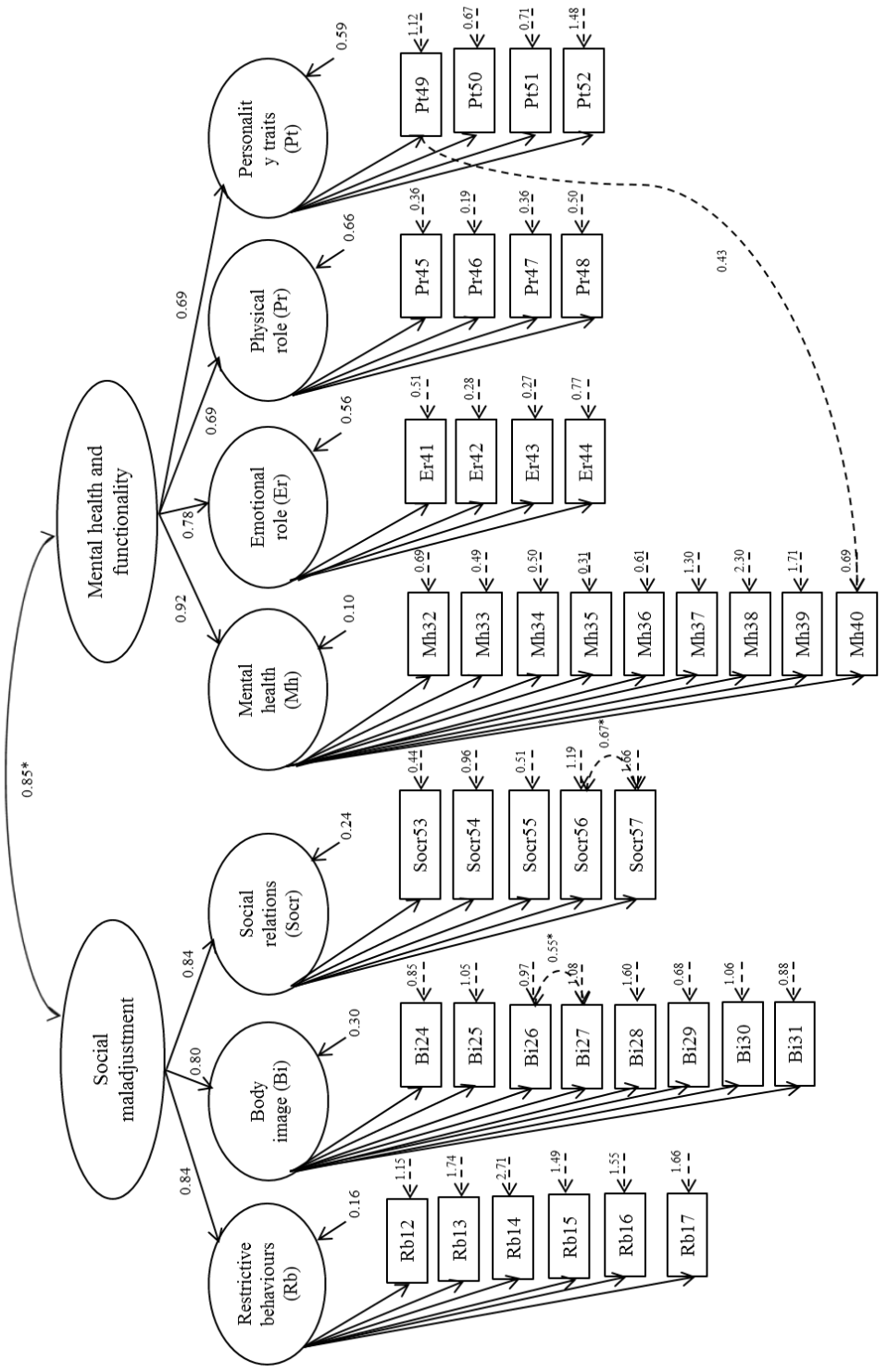


Figure 6.1. Path diagram of the resulting structure of the HeRQoLEDv2. In order to keep the path diagram from becoming overly complex, the lowest and highest factor loadings for each domain are described here: Restrictive behaviors = 0.49 - 0.71, Body Image = 0.70 - 0.87, Social relations = 0.57 - 0.89, Mental Health = 0.54 - 0.85, Emotional role = 0.81 - 0.94, Physical role = 0.84 - 0.95 and Personality Traits = 0.64 - 0.84. * Indicates covariances among exogenous variables.

Rasch analysis to obtain the shortened version

Data from all 324 ED patients who responded at T1 were used for the Rasch Rating Scale analysis.

Originally, the social maladjustment domain was composed of 19 items. Nine of them were removed because they showed inadequate fit indices (infit or outfit) or because they overlapped the same level of difficulty as other items. Experts in ED evaluated the importance of the item content before removing the item. The shortened social maladjustment domain consisted of 10 items separated by 0.10 or more logit values. Table 6.1 shows the characteristics of the measurement level, standard error, infit, outfit, and item total correlations. The level of difficulty is represented by the trait level (δ), where high values indicate greater difficulty with social adjustment.

Four items of the social maladjustment domain did not comply with all the requirements for adequate functioning of rating scale categories. Specifically, fewer than 10 participants had endorsed the response category "Almost always" in the item RB12 "Do you fast for a day although you feel hungry". We combined adjacent categories "almost always" with respondents of "Always" to obtain a robust structure of high frequency categories. This combination reproduced satisfactory results with an outfit index of 1.3. Items RB15 "Do you avoid eating with others?" and BI27 "Do you worry about the possibility of gaining weight?" showed large outfits in one of their response categories. Response category "Always" of item RB15 presented an outfit index of 2.1. After combining respondents of adjacent categories "always" and "almost always", the outfit index reduced to 1.5. For the item BI27 the category response "never" presented an outfit index of 2.6. After combining this response category with the adjacent category of "almost never" the outfit value reduced to 1.4. The fourth problematic item is SOCR54 "Do you think that your eating habits negatively affect your family relationship?" which presented a large outfit (Outfit = 1.9) for response category 3 "several times" but not for the remaining of the response categories.

Table 6.1. Rasch model: Item measure, SE, fit statistics and item-total correlations of the social maladjustment domain.

		Social maladjustment				
Item[a]	Content	δ	SE	Infit	Outfit	r_t
(1) RB12	Do you fast for a day, although you feel hungry?	1.54	0.07	1.17	0.92	0.58
(2) RB13	Do you skip some meals, although you feel hungry?	0.56	0.05	1.33	1.12	0.69
(3) RB15	Do you avoid eating with others?	0.48	0.05	1.12	1.19	0.61
(10) SOCR56	Do you think that your eating habits negatively affect your personal relationship or the possibility of finding one?	0.23	0.05	1.21	1.16	0.63
(9) SOCR54	To what extent do your concerns about eating negatively affect your family relationship (talking less, discussing more, diminished confidence?)	0.12	0.05	1.03	1.17	0.62
(8) BI28	Do you avoid situations in which others can see your body, for example, in the gym, the pool, or on the beach?	-0.01	0.05	1.26	1.23	0.70
(4) BI24	In general, do you feel fat, despite the fact that other people (family, friends, doctors, etc.) tell you otherwise?	-0.40	0.05	0.83	0.80	0.81
(5) BI25	Do you think that some parts of your body, for example, hips, waist or thighs, are too big or wide compared with the rest of your body?	-0.52	0.05	0.95	0.89	0.78
(6) BI26	Do you worry about your weight?	-0.92	0.05	0.82	0.80	0.79
(7) BI27	Do you worry about possibly gaining weight?	-1.09	0.06	0.77	0.84	0.78

Every question has a response scale of 6 ordinal options, being 0 = Never and 5 = Always.

δ = Level of severity of the social maladjustment factor. Higher values indicate higher severity; SE = standard error; r_t = correlation between item and total measured social maladjustment level based on the Rasch calibrated item scores and total scores.

[a] The numbers in parentheses reflect the current item location in the shortened version.

This English translation has not been validated linguistically. We provide an approximate translation of the Spanish items into English.

Combining adjacent categories was not a good approach since the resulting merged response category would count with an excessive number of respondents. We decided to leave the item as it was.

Unidimensionality was supported since the PCA of the residuals did not give additional factors with eigenvalues exceeding 3.00. Furthermore, the fit indices ranged from 0.77 to 1.30. All the item total correlations were high and homogeneous (see Table 6.1). Differential item functioning was observed only in one item, BI27 "Do you worry about the possibility of gaining weight?" with a difference slightly higher than 0.5 (DIF contrast = 0.66; $p < 0.05$) between the anorexia and the bulimia subgroups. For patients with anorexia nervosa, this item was slightly more difficult than for patients with bulimia. Intercorrelations between residuals were all below 0.50 (range -0.30 to 0.47).

The final shortened scale of social maladjustment included 10 items. The item locations for the HeRQoLED-S are shown in Figure 6.2 (left-hand side). The person separation index (2.46) and the item separation index (12.48) exceeded the required value of 2.0, thereby indicating a reliability above 0.80. The total score was transformed to range from 0 to 100 (mean score: 48.8; SD: 23.2).

The mental health and functionality scale originally included 21 items. After performing iterative Rasch analyses and item content analysis, 11 of them were removed because they overlapped or misfit and were not clinically essential. Seven of the 10 remaining items in the scale were separated by 0.10 logit units and 3 of which were separated by 0.04 logit units (Table 6.2; Figure 6.2, right). The 3 overlapping items (Figure 6.2, right-hand side) were retained because they were considered clinically meaningful based on expert opinion and had adequate fit indices.

Unidimensionality was supported since the PCA of the residuals did not lead to additional factors with eigenvalues exceeding 3.00. Furthermore, the fit indices ranged from 0.72 to 1.27. The item total correlations were all high and homogeneous, ranging from 0.61 to 0.78.

Only two items of the mental health and functionality domain did not comply with the requirements for adequate functioning of rating scale categories. Specifically,

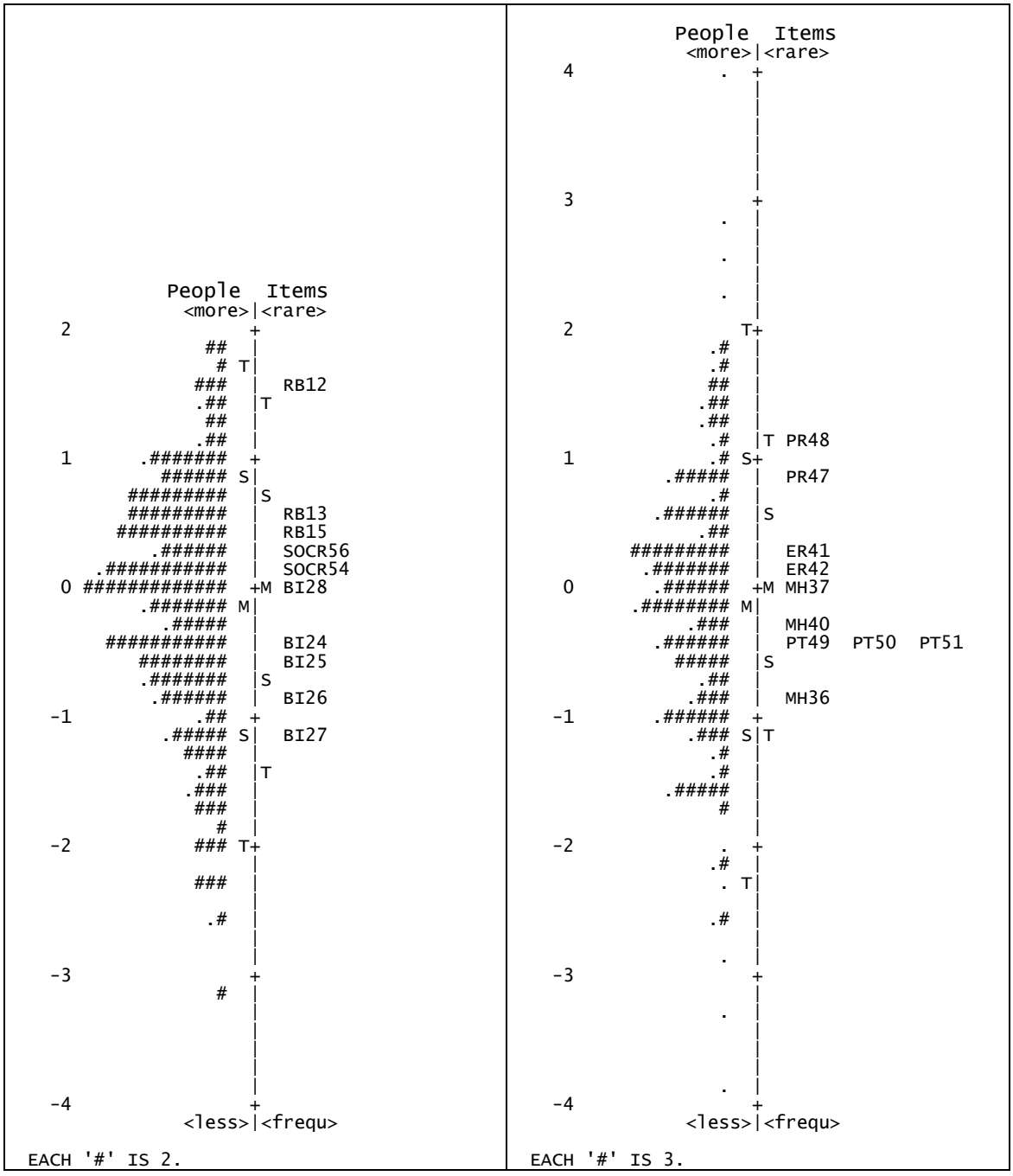


Figure 6.2. Person and item map of the social maladjustment and mental health and functionality domains. Both individuals and items are presented in the same logit scale. Social maladjustment items are presented on the left side, and mental health and functionality items are on the right side. Items are summarized by the acronym of their corresponding first-order factor along with the number they had in the original HeRQoLEDv2. Tables 1 and 2 present a brief description of each question's content. RB = restrictive behaviors; SOCR = social relations; BI = body image; PR = physical role; ER = emotional role; MH = mental health; PT = personality traits.

Table 6.2. Rasch model: Item measure, SE, fit statistics and item-total correlations of the mental health and functionality domain

		Mental health and functionality				
Item[a]	Content	δ	SE	Infit	Outfit	r_t
(7) PR48	Do you have to stop performing some tasks as a result of your physical problem?	1.16	0.06	1.04	1.05	0.67
(6) PR47	Do you find it difficult to maintain attention as a result of your physical problem?	0.80	0.06	1.11	1.03	0.71
(4) ER41	Do you have to make an extra effort or invest more time than usual as a result of your emotional problems?	0.27	0.06	0.86	0.81	0.76
(5) ER42	Do you accomplish less than you would like to as a result of your emotional problem?	0.17	0.06	0.90	0.85	0.77
(2) MH37	Do you have very sudden mood changes that you find difficult to control?	0.06	0.06	1.17	1.20	0.62
(3) MH40	Do you feel worthless?	-0.35	0.06	0.73	0.73	0.80
(10) PT51	Do you set very high goals and feel dissatisfied if you do not meet them?	-0.39	0.06	1.17	1.14	0.68
(9) PT50	Do you think that you have to do things perfectly or just not to do them at all?	-0.42	0.06	1.24	1.24	0.68
(8) PT49	Do you feel lack of confidence in your own capabilities?	-0.45	0.06	0.87	0.88	0.73
(1) MH36	Do you feel happy?	-0.86	0.06	0.84	0.95	0.64

Every question has a response scale of 6 ordinal options, being 0 = Never and 5 = Always.

δ = Level of severity of the social maladjustment factor. Higher values indicate higher severity; SE = standard error; r_t = correlation between item and total measured social maladjustment level based on the Rasch calibrated item scores and total scores.

[a]The numbers in parentheses reflect the current item location in the shortened version.

This English translation has not been validated linguistically. We provide an approximate translation of the Spanish items into English.

the category response "Always" of item PR48 "Do you have to stop performing some tasks as a result of your physical problem?" presented an outfit index of 2.2. Therefore, we decided to combine this response option with the adjacent category "Almost always". After this combination, the outfit reduced to 1.4. The category response "Never" from the item MH36 "Do you feel happy?" was only reported by 1 participant. Thus, we decided to combine it with the adjacent category response "Almost never" to enlarge the sample. After this combination, the outfit index was -1.58.

Figure 6.2 (right side) shows the item and person locations along the logit scale. Positive values indicate high levels of mental health disease and dysfunction, whereas negative values indicate low levels of mental health disease and dysfunction.

The person separation index (2.5) and the item separation index (9.7) for this sample also exceeded the required value of 2.00, indicating a reliability of the scale above 0.80. The raw score in this domain was also transformed to range from 0 to 100 (mean = 48; SD = 20.3). Statistically significant DIF contrasts were not observed for any item of the scale.

Intercorrelations between residuals were below 0.50 (range -0.29 to 0.41), except for two items ("Do you have to stop performing some tasks as a result of your physical problem?" and "Do you find it difficult to maintain the attention as a result of your physical problem?") which slightly surpassed this threshold (r = 0.51).

In summary, after applying the Rasch rating scale analysis to the original 40 items (after excluding items from binges and symptoms domains) of the HeRQoLEDv2 we obtained a shortened version of 20 questions divided in 2 factors, 'social maladjustment' and 'mental health and functionality'. This HeRQoLED short version provides separate scores for each factor. Calculating the score in both long and short versions requires summing the response options selected in the factor's items, standardizing the score to range from 0 to 100. In case of missing values we applied the mean imputation method.

Confirmatory Factor Analysis of the HeRQoLED-S

Data from the 207 patients who returned questionnaires at T2 without missing answers were used for the CFA of the HeRQoLED-S. The hypothesized model was similar to that of the long version but included only the 20 items accepted after the Rasch analysis. The Lagrange multiplier test was again used. The first pair of errors

intercorrelated belonged to two items from the body image domain, and the second to the personality traits domain.

The factorial structure that resulted after allowing for these covariances between errors proved satisfactory since it resulted in acceptable fit indices (χ^2 (df = 160) = 305.96, $p < 0.0001$; χ^2/df ratio = 1.9; RMSEA = 0.07; NNFI = 0.93 and CFI = 0.94) and significant factor loadings (Figure 6.3).

Concurrent validity and reliability of the HeRQoLED-S

A data set for the HeRQoLED-S was created using the responses of all 245 patients who completed questionnaires at T2. As hypothesized (Table 6.3) the social maladjustment factor correlated more strongly with the dieting concern factor of the EAT-26 ($r = 0.82$, $p < 0.001$) than with the remaining factors. The mental health and functionality factor of the HeRQoLED-S also correlated higher with the mental summary component of the SF-12 ($r = -0.69$, $p < 0.0001$) than with the other factors. The Cronbach alpha was 0.91 for the social maladjustment domain and 0.90 for the mental health and functionality domain.

6.5. Discussion

This study confirmed the internal structure of a newly developed questionnaire for eating disorders, the 55- question HeRQoLEDv2. We also applied CFA and Rasch analysis to develop a shorter 20-question version, which maintained satisfactory psychometric qualities, and we validated the internal structure of the shortened questionnaire.

Of the three other disease-specific instruments created to date for measuring HRQoL in patients with an ED, only the EDQOL questionnaire (Engel et al., 2006) has been subjected to a CFA. In that study, the investigators confirmed the structure of a second-order factor, presumed to be the HRQoL construct that explained the relationships between four latent first-order factors. In the current

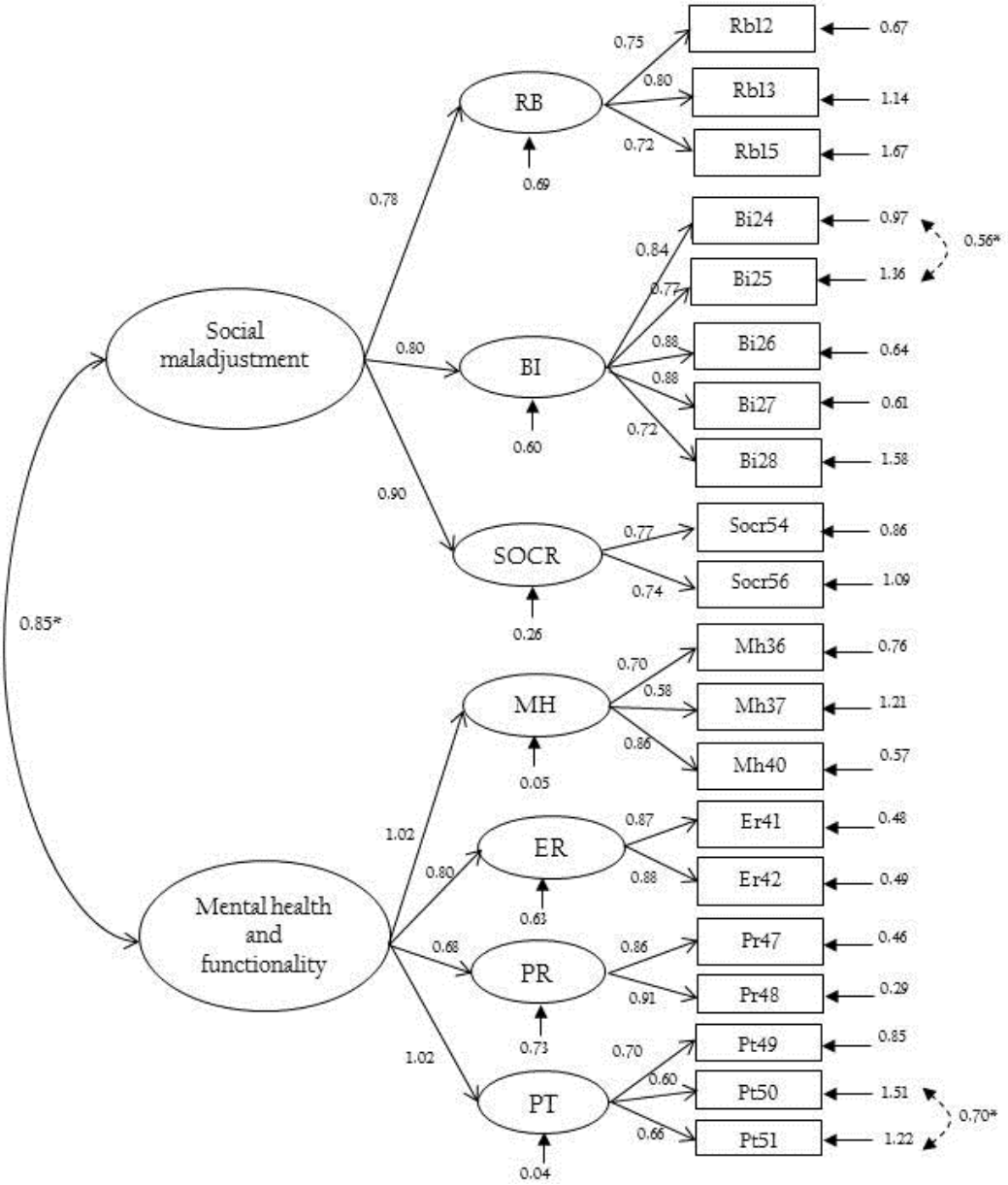


Figure 6.3. Confirmed factor structure of the HeRQoLED-S. RB = restrictive behaviors; BI = body image; SOCR = social relations; MH = mental health; ER = emotional role; PR = physical role; PT = personality traits. * Indicates covariance among exogenous variables.

study, CFA of the HeRQoLEDv2 revealed two correlated second-order factors that explained the relationships between seven first-order factors. In theorizing our model, we did not hypothesize an orthogonal structure a priori because we assumed that the HRQoL measurement construct included the intercorrelation of physical, mental, and social factors affected by EDs and treatment (Fayers and Machin, 2000; Revicki et al., 2000).

Table 6.3. Measurement of the concurrent validity and reliability of the HeRQoLED-S

	Social maladjustment	Mental health and functionality
SF-12 PCS	-0.27	-0.33
SF-12 MCS	-0.51	-0.69
EAT "Dieting concern" factor	0.82	0.61
EAT "Bulimia and food preoccupation" factor	0.73	0.58
EAT "Oral control" Factor	0.49	0.44
Cronbach alpha	0.91	0.90

All correlations were assessed using Pearson correlation coefficient. * All correlations were statistically significant at $p < 0.0001$; SF-12 PCS = Short-Form-12, Physical Component Summary; SF-12 MCS = Short-Form-12, Mental Component Summary

Validation of the HeRQoLEDv2 using CFA provides the questionnaire with greater construct validity than in the version we previously developed (Las Hayas et al., 2007). To perform the CFA, we recruited 262 patients with ED. Although one could argue that this sample size is small considering the length of the questionnaire, it must be noted that it is difficult to recruit patients with ED, so recruiting this amount of participants can be considered as strength of the study more than a limitation. Among the potential statistical drawbacks derived from the sample size are the increase in sampling error, instability, and reduced reliability of factor analysis solutions (MacCallum et al., 1999).

A second aim of this study was to use modern analytical techniques to create a shorter version of the HeRQoLEDv2. Various strategies are available for the reduction of questionnaires (Beaton et al., 2005). We chose to apply the Rasch method, as this technique produces a scale that calibrates items based on their range of difficulty for the target population.

The 20-item HeRQoLED-S that emerged from the Rasch method provided adjustment levels (infit and outfit), unidimensionality, and local independence sufficient to be considered adequate. A slight DIF was observed in only one item. We decided not to remove the item from the questionnaire since it was clinically

relevant and presented satisfactory levels of functioning in the other parameters (fit statistics, local dependence, and response scale functioning).

A third aim of the study was to validate the HeRQoLED-S. A CFA applied to the HeRQoLED-S confirmed the goodness of the structure achieved using the Rasch method, as reflected in the obtained fit indices. Other studies have also applied CFA to validate the internal structure of shortened questionnaires (Calvete et al., 2005). Apart of the hypothesized concurrent correlations between the HeRQoLED-S and specific domains of the EAT-26 and SF-12, the social maladjustment factor of the HeRQoLED-S correlated highly with the second factor of the EAT-26, bulimia and food preoccupation. This latter correlation had not been hypothesized previously. The bulimia and food preoccupation factor contains questions about the control that food exercises over an individual's life and about binges and vomiting. It makes sense that the social maladjustment domain is highly correlated with this factor because individuals who engage in bingeing and vomiting also manifest problems with social adjustment (Beales and Dolton, 2000; Rorty et al., 1999). However, our first hypothesis was to correlate the social maladjustment domain with the dieting concern factor because questions pertaining to it inquire about restrictive behaviors and body image, and are more similar in content to those covered by the social maladjustment domain.

We estimated that the shortened form requires approximately 5 to 7 minutes to complete, which is about one-third the time it takes to complete the original HeRQoLEDv2. This is a considerable reduction in time commitment for participants.

One limitation of the HeRQoLED-S is that its items did not cover the entire range of existing difficulties, and gaps in construct difficulty were detected. Although including more items would have helped cover the different levels of construct difficulty, this was not possible because we were working with a predetermined set of items and selected those that provided the best distribution despite the gaps. In addition, although redundant items were identified for mental health and

functionality, they were maintained because the scale generally provided good content validity and good fit indices.

Coste et al. (1997) have recommended that shortened versions of questionnaires be evaluated psychometrically (particularly with regard to construct validity and reliability) using a new and independent sample. Due to financial limitations and difficulties in recruiting another large sample of patients with an ED, the HeRQoLED-S was validated using the same patient sample as in the follow-up study. We believe this was appropriate given that the T2 sample contained a different number of patients and that the one-year interval since the last contact uses to lead to significant changes in ED symptoms, as some other studies have shown (Aranda et al., 1997; Bowers and Ansher, 2008). Nevertheless, the same level of validity cannot be obtained from a repeat sample as from a new independent sample. Thus, the shortened HeRQoLED-S must still be validated among different groups of patients with eating disorders.

In conclusion, CFA analysis supports an internal structure of two latent factors of the 55-question HeRQoLEDv2. A short form questionnaire derived from this second order structure, the 20-item HeRQoLED-S, has been developed and validated with modern psychometric techniques that facilitate its use in research and clinical practice. Both versions have demonstrated good reliability and validity. Future applications of HeRQoLEDv2 and HeRQoLED-S using different ED patient samples will yield more evidence about their validity and reliability.

Capítulo 7

Cross-validation study using Item Response Theory: the Health-Related Quality of Life for Eating Disorders questionnaire-Short Version

Este capítulo ha sido publicado como investigación original en la revista Assessment (Factor de Impacto: 3,286; 1er cuartil), con referencia: Amaia Bilbao, Carlota Las Hayas, Carlos G Forero, Ángel Padierna, Josune Martin, José M Quintana. Cross-validation study using item response theory: the Health-Related Quality of Life for Eating Disorders questionnaire-short version. Assessment 2013;21(4):477-493.

7.1. Abstract

The Health Related Quality of Life for Eating Disorder Short questionnaire is one of the most suitable existing instruments for measuring quality of life in patients with eating disorders. The objective of the study was to evaluate its reliability, validity and responsiveness in a cohort of 377 patients. A comprehensive validation process was performed, including confirmatory factor analysis (CFA) and a graded response model (GRM), and assessments of reliability and responsiveness at one year of follow-up. The CFA confirmed the two second-order latent traits, social maladjustment, and mental health and functionality. The GRM results showed that all items were good for discriminating their respective latent traits. Cronbach's alpha coefficients were high, and responsiveness parameters showed moderate changes. In conclusion, this short questionnaire has good psychometric properties. Its simplicity and ease of application further enhance its acceptability and usefulness in clinical research and trials, as well as in routine practice.

7.2. Introduction

The incidence of eating disorders (EDs) has increased in recent years, especially among women, affecting millions of people worldwide (Gonzalez et al., 2001; Mitchison et al., 2012; Steinhausen, 2002; van Hoeken et al., 1998). EDs are diseases that primarily affect young women; have a great impact that is manifested in terms of physical, psychological, and social functioning problems; and have a significant tendency to become chronic, with many patients only partially recovering (Gonzalez et al., 2001; Hsu, 1995; Hudson, et al., 2007; Steinhausen, 1995 and 2002).

In recent years, a growing importance is being given to the evaluation of patient health-related quality of life (HRQoL) as a measure of health status as referred by them and as a tool to evaluate the results of interventions, especially with regard to chronic diseases (Boini et al., 2004; Gonzalez et al., 2001). In the case of

psychiatric disorders, assessment of HRQoL can serve to complement the clinical perspective, providing information about the disease and its impact on the lives of patients. More specifically, over the past decade, the interest in HRQoL of patients with ED has dramatically increased (Bamford, 2010; Basu, 2004; Jenkins et al., 2011). Previous studies have shown that these disorders do indeed have a major impact on several HRQoL areas (de la Rie et al., 2005 and 2006; Jenkins et al., 2011; Mitchison et al., 2012; Padierna et al., 2000 and 2002; Spitzer et al., 1995; Wade et al., 2012), having a strong and consistently negative impact on HRQoL, compared with assessments in people without EDs (Bamford, 2010; de la Rie et al., 2006; Padierna et al., 2000; Wade et al., 2012). However, most of such studies used generic instruments to assess the impact of EDs on physical, mental, and social factors (Hay and Mond, 2005; Las Hayas et al., 2010; Mitchison et al., 2012; Wade et al., 2012), which is a weakness because generic questionnaires may not capture the magnitude of the disability caused by the illness (Keilen et al., 1994), and their content may not be relevant in the context of EDs (Guyatt et al., 1993; Marquis et al., 2004). In contrast, specific questionnaires have greater discriminatory power for identifying the severity and response to treatment of the disease (Engel et al., 2009; Las Hayas et al., 2006; Wiebe et al., 2003).

Furthermore, a major use of health measurement scales is to monitor changes in health status over time, and efficiency may be a priority, that is, responses should be obtained using the shortest possible questionnaire (Moran et al., 2001). One of the major disadvantages of self-administered questionnaires is always the burden of their completion (Aaronson et al., 2002). Another limitation of self-administered questionnaires is the low response rate, which greatly affects the validity of a study. It has, however, been shown that the use of short questionnaires significantly increases the response rate and the compliance (Kalantar and Talley, 1999). In clinical practice, where information is collected to assess response to treatments, the goal is always to involve as little effort as possible for both the patient and the physician. Therefore, a reduced version would be more useful in epidemiological studies, clinical trials, and clinical practice, since short

questionnaires improve compliance of patients and response rates and are designed to improve the quality of response (Coste et al., 1997).

The first specific HRQoL questionnaires for patients with EDs were created almost simultaneously (Abraham et al., 2006; Adair et al., 2007; Engel et al., 2006; Las Hayas et al., 2006). As stated by Engel et al. (2009), although these instruments have some similarities, they vary in length, subscales, or domains assessed, methods of development, and characteristics of the validation sample. Furthermore, as stated by Adair et al. (2007), two of these questionnaires focus predominantly on symptoms and behaviors (Abraham et al., 2006; Las Hayas et al., 2006), one was tested on inpatients only (Abraham et al., 2006), one addresses four broader domains but has no ED symptom and behavior items (Engel et al., 2006), and one was developed considering its suitability for adolescents (Adair et al., 2007). Tirico et al. (2010) conclude that only three of these instruments (Adair et al., 2007; Engel et al., 2006; Las Hayas et al., 2006 and 2007) were adequate in terms of development and validation, and in addition, they also conclude that the Health-Related Quality of Life in Eating Disorders (HeRQoLED), developed by our research group (Las Hayas et al., 2006 and 2007), was the one that presented the best psychometric properties, and the only instrument for which responsiveness has been reported, presenting satisfactory results.

The HeRQoLED is a specific self-administered questionnaire consisting of 55 items that cover 9 domains: mental health, emotional role, physical role, personality traits, social relations, body image, restrictive behaviors, binges, and symptoms (Las Hayas et al., 2006). We performed an exhaustive psychometric study, with a sample of 324 ED patients recruited during 2002, demonstrating that the questionnaire was valid, reliable, and sensitive to change (Las Hayas et al., 2006 and 2007). However, a limitation of this questionnaire is that it takes a long time to complete due to its large number of items. Therefore, we decided to develop a short form of the HeRQoLED questionnaire using Rasch analysis. For this reduction, only seven of the original nine domains were considered, excluding the symptom domain, because it is a list of symptoms rather than a measurement

scale, and the binge domain, since it is a separate issue that only affects people who binge. Accordingly, with the 40 items composing the seven other factors, first the internal structure of the questionnaire was confirmed showing two second-order latent traits: "social maladjustment" (SocM) and "mental health and functionality" (MHF). These summary measures were subjected to Rasch analysis, and finally the HeRQoLED short form (HeRQoLED-S) was derived comprising 20 items (Las Hayas et al., 2010). Both the confirmation of the second-order internal structure and the reduction of the HeRQoLED questionnaire were carried out with the same sample of 324 ED patients used for the development and validation processes of the original questionnaire.

A first validity study of the HeRQoLED-S was performed using the follow-up sample from the original sample 1 year after initial assessment, and this indicated that thereduced questionnaire was valid and reliable (Las Hayas et al., 2010); however, this first validation of the HeRQoLED-S was limited in that the same sample had been used both for the development as for the confirmation of its structure. It is recommended to use new and independent samples for psychometric studies of short forms of questionnaires (in particular regarding the construct validity and reliability; Coste et al., 1997), but in our case this was not possible given financial limitations and the difficulty of obtaining large samples in these patients, and consequently, we used the follow-up sample. Although we believe that this was reasonable, given that the 1 year interval since the previous contact meant that there were significant changes in ED symptoms, we understand that this approach does not obtain the same level of validity as we would have achieved if we had used an independent sample. Furthermore, in this first validity study, the construct validity, which refers to the examination of the theoretical relationship of the items to each other and to the hypothesized scales (Fayers and Machin, 2007), was only studied from the perspective of classical test theory, being another limitation. Another weakness of that study, as noted by other authors (Bamford, 2010), was the lack of an analysis of the responsiveness of the HeRQoLED-S, which refers to the ability of an instrument to detect changes when a patient improves or deteriorates (Fayers and Machin, 2007).

Given the importance of measuring HRQoL in patients with EDs, more research being needed in the field (Bamford, 2010), it is essential to have valid and reliable specific questionnaires to ensure that they are adequate and appropriate. As some authors have pointed out (Bamford, 2010), the HeRQoLED-S is a useful contribution in this field, since it is one of the most suitable existing instruments for measuring HRQoL in patients with ED. Consequently, it is all the more important to assess its validity, reliability, and responsiveness more thoroughly.

The objective of the present study was to conduct a validation process of the short form of the questionnaire, the HeRQoLED-S, in a new cohort of patients, confirming the second-order internal structure, and examining its properties using item response theory (IRT) for greater accuracy in the estimation of parameters characterizing the instrument. Furthermore, other psychometric properties such as known-groups validity, convergent and discriminant validity, reliability, and responsiveness were also analyzed. Regarding convergent validity, we hypothesized that the SocM and MHF scales would correlate more strongly with the dieting concern factor of the Eating Attitudes Test-26 (EAT-26; Garner et al., 1982), and the mental component summary (MCS) domain of the 12-item Short Form Health Survey (SF-12; Gandek et al., 1998), respectively. In relation to known-groups validity, we hypothesized that the more severe the ED, the higher the score on the EAT-26, and the lower the score on the MCS domain of the SF-12, the higher patient HeRQoLED-S scores would be. Overall, the goal of the study was to perform a comprehensive assessment of the reliability, validity, and responsiveness of the short form of the questionnaire using a new sample of patients, combining classical and modern techniques, and increasing the methodological rigor.

7.3. Methods

7.3.1. Participants

The current study included data concerning clinical cases of ED from two prospective cohorts recruited independently by different public mental health services in the province of Biscay, Spain. Given that recruiting a large clinical sample in a disorder with a low incidence is very difficult, we decided to combine two different samples recruited from two independent studies carried out by our research group. The first sample was recruited during 2003 and the second during 2007. In both cohorts, participants consisted of all patients with an ED who were treated (for their ED) by any of our collaborating psychiatrists in the established periods, and who fulfilled the following criteria for eligibility: (a) having been diagnosed with an ED according to the *Diagnostic and Statistical Manual for Mental Disorders-IV* (American Psychiatric Association, 1994) by any of our psychiatrists; (b) having received outpatient treatment by the time of the study in any of the three mental health services that collaborated in this study; (c) being free from any clinically significant multiple organic disorders, cerebral organic deterioration, acute psychosis, or languages barriers that would prevent patients from completing the questionnaires; and (d) agreeing to participate voluntarily after being provided with information about the study personally by his or her psychiatrist and providing informed consent. Each hospital's ethics review board approved the study, and the tenets of the Declaration of Helsinki were followed. Furthermore, we did estimate the sample size needed for an accurate study of validity of our questionnaire. Based on the number of items of the HeRQoLED-S, a sample of 200 participants, 10 participants per item following the criterion established by Nunnally (1978), would suffice.

From the 324 and 169 patients included in the first and second cohorts, respectively, who fulfilled selection criteria, 245 (75.62%) and 132 (78.11%), respectively, completed the questionnaires and gave written informed consent.

The mean age of the patients with an ED was 27.26 years (*SD* = 8.85), and the great majority of them were women (96.29%; see Table 7.1).

Table 7.1. Characteristics of the patients of each cohort (n = 377).

Variables	Total (n = 377)	Cohort 1 (n = 245)	Cohort 2 (n = 132)
Age, $\bar{x}$ (SD)	27.26 (8.85)	28.09 (8.76)	25.70 (8.84)
Women, n (%)	363 (96.29)	233 (95.10)	130 (98.48)
Severity Index*, n (%)			
Minor	119 (35.31)	82 (40)	37 (28.03)
Moderate	108 (32.05)	62 (30.24)	46 (34.85)
Severe	110 (32.64)	61 (29.76)	49 (37.12)
Diagnosis, n (%)			
Anorexia nervosa	109 (28.91)	52 (21.22)	57 (43.18)
Bulimia nervosa	64 (16.98)	32 (13.06)	32 (24.24)
EDNOS	204 (54.11)	161 (65.71)	43 (32.58)
HeRQoLED-S, $\bar{x}$ (SD)			
SocM	47.44 (24.23)	44.61 (24.64)	52.74 (22.61)
MHF	45.02 (21.24)	43.35 (20.35)	48.15 (22.57)
SF-12, $\bar{x}$ (SD)			
MCS	37.14 (11.87)	38.71 (11.55)	34.76 (12)
PCS	48.97 (9.31)	50.23 (9.09)	47.07 (9.36)
EAT-26, $\bar{x}$ (SD)			
Dieting	14.26 (10.73)	13.08 (10.36)	16.43 (11.08)
Bulimia and food preoccupation	5.76 (4.83)	5.12 (4.84)	6.93 (4.61)
Oral control	5.53 (5.20)	4.73 (4.95)	7.01 (5.34)
Total score	25.55 (17.77)	22.94 (17.66)	30.37 (17.01)

Data are given as number (percentage) for categorical variables and as mean (standard deviation) for continuous variables.

*Severity of illness if defined by means of the Clinical Global Index (CGI), which consists of a single question with a five-point scale, from 0 (not severe) to 5 (very severe). Patients were classified as "minor" if he or she received a score of 0, "moderate" if the patient had a score of 1 or 2, and "severe" with a score of 3 or 4.

SD: Standard deviation; EDNOS: Eating disorder not otherwise specified; HeRQoLED-S: Health-Related Quality of Life in Eating Disorders short form; SocM: Social maladjustment; MHF: Mental health and functionality; SF-12: 12-item Short Form Health Survey; MCS: Mental component summary; PCS: Physical component summary; EAT-26: Eating Attitudes Test-26.

7.3.2. Measure

Each participant provided sociodemographic data and completed the following three assessment measures.

The HeRQoLED-S, the focus of this study, is a specific self-administered questionnaire, developed in Spanish, which is intended to assess the impact of EDs on patient physical, psychological, and social functioning (Las Hayas et al., 2010). It consists of 20 items covering two areas: SocM and MHF. The SocM domain is composed of items regarding restrictive behaviors, body image, and social relations, which measure the impact of having an ED on the sociocultural life. The MHF domain is composed of items that measure the impact of having an ED on the physical and emotional role, mental health, and personality traits. The scores for these areas are obtained by summing the items in each corresponding scale and standardizing to range from 0 to 100, where a higher score indicates poorer HRQoL.

The SF-12 (Gandek et al., 1998) is a short generic survey of HRQoL and its results can be summarized by two subscores: the Physical Component Summary (PCS) and the Mental Component Summary (MCS). These scores also range from 0 to 100, with higher values indicating better health perception. Its validity in Spanish patients has been demonstrated (Alonso et al., 1995).

The EAT-26 (Garner et al., 1982), which assesses the behavioral and cognitive characteristics of ED patients, consists of three scales, dieting, bulimia and food preoccupation, and oral control, where the higher the score, the greater the level of ED symptomatology. It also provides a total score of between 0 and 76 with a cutoff value of 20. Scores above 20 indicate the presence of behaviors or thoughts characteristic of ED individuals. It has also been validated in a Spanish population (Castro et al., 1991).

Furthermore, the psychiatrists who collaborated in the studies completed a short clinical questionnaire, in which they provided information about the clinical diagnosis of the patient according to *DSM-IV* diagnostic criteria (American Psychiatric Association, 1994) and the Clinical Global Index (CGI; Guy, 1976). This consisted of a single question with a 5-point scale, on which psychiatrists had to grade the severity of the patient's condition from 0 (*not severe*) to 5 (*very severe*).

7.3.3. Procedure

All the measurement instruments were posted to the participants for them to complete at home. Two reminders also were sent at fortnightly intervals to those who did not respond to the first mailing. The first reminder was a letter in which patients again were requested to complete the measurement instruments that had been sent previously, while the second reminder included all the measurement instruments, in addition to a letter. The collaborating psychiatrists completed the clinical questionnaire and collected the sociodemographic data.

Patients from the second cohort were followed-up 1 year later to study the responsiveness of the HeRQoLED-S. For the follow-up study, the test battery was again posted to patients, as were the reminders for those not responding as described previously.

7.3.4. Statistical Procedures

To describe the samples, we used means and *SD*s, frequencies, and percentages.

Regarding construct validity, first, the two-factor second-order internal structure of the HeRQoLED-S was tested using confirmatory factor analysis (CFA) for categorical variables. The structural and measurement model hypothesized for the short form was the same as that of the long version, the only difference being that fewer items were assigned to each first-order factor (Las Hayas et al., 2010). Besides, the one-factor second-order structure was also examined. Then, we

extended the psychometric evaluation of the construct validity using IRT, which unlike classical test theory, focuses on item-level properties rather than on scale-level properties (Gomez, 2011). Specifically, the graded response model (GRM) was used, which is useful in the evaluation of polytomous items (Samejima, 1969). Convergent validity was assessed by analyzing the relationship between the scores on the HeRQoLED-S domains and the SF-12 and EAT-26 domains with the Pearson correlation coefficient, and following the same hypothesis used previously (Las Hayas et al., 2010). Known-groups validation was examined by comparing the HeRQoLED-S mean scores among the different groups according to the CGI, their total score on the EAT-26, and the MCS domain of the SF-12. Furthermore, the magnitude of group differences was estimated by means of the effect sizes (ES).

Regarding reliability, internal consistency was assessed using Cronbach's alpha coefficient (Cronbach, 1951).

To study responsiveness only the second cohort was used, and patients were classified as "improved" or "worsened". As we did not have a transition question, the classification of patients was conducted based on the change scores of the total EAT-26 score. We used the EAT-26 instead of the SF-12 instrument because, in general, the EAT-26 correlates more strongly with the HeRQoLED-S (Las Hayas et al., 2010), and it is more specific instrument for ED than the SF-12. Ceiling and floor effects and means and *SD*s were calculated for the HeRQoLED-S scales in each group of patients at baseline and after 1 year. To measure the responsiveness of the HeRQoLED-S, we used the standardized effect size (SES) and standardized response mean (SRM) (Guillemin et al., 1993).

All effects were considered statistically significant at $p < 0.05$. All statistical analyses were performed with SAS for Windows statistical software (Version 9.2; SAS Institute, Cary, NC), except for the CFA and GRM analyses wherein we used the Mplus software (Version 6.1; Muthén and Muthén, 2010).

7.4. Results

Construct validity

The results of the CFA for the hypothesized second-order model of two latent factors, SocM and MHF, are shown in Figure 7.1, for which an unweighted least square mean and variance adjusted estimator was used. Specifically, the second-order factor SocM is affected by the "restrictive behaviors," "body image," and "social relations" first-order factors, and the MHF second-order factor is affected by the "physical role," "emotional role," "mental health," and "personality traits" first-order factors. Several fit indices were calculated (Hu and Bentler, 1999; Shevlin and Miles, 1998; Mulaik et al., 1989; Muthén and Muthén, 2010): the chi-square statistic with degrees of freedom; the root mean squared error of approximation (RMSEA), for which a value below .06 was considered acceptable; and the Tucker–Lewis index (TLI) and comparative fit index (CFI), both of which had to exceed 0.95 to be satisfactory. The results of the CFA provided satisfactory fit indices (Figure 7.1): RMSEA = 0.066, CFI = 0.97, TLI = 0.96, and χ^2 = 398.48 (df = 162). All factor loadings were significant ($p < 0.001$) with a range from 0.61 to 0.93, and the correlation coefficient between the two factors was high (r = 0.86).

The results of the CFA for the one-factor second-order model were as follows: RMSEA = 0.075, CFI = 0.95, TLI = 0.95, and χ^2 = 474.95 (df = 163). Comparing the fit indices of one-factor and two-factor second-order structures, those obtained from one-factor second-order structure were slightly worse than those obtained in the two-factor structure (Figure 7.1). This fact is confirmed by higher value of chi-square, slightly lower CFI and TLI indices, and higher RMSEA value in the one-factor model than in the two-factor model. For the comparison of both models, as the models are nonnested, we used the Akaike information criterion (AIC). The AIC is calculated from the value of the chi square through penalizing the sum of the parameters (Raftery, 1995), where lower values indicate that the model fits better. In the two-factor second-order model, the AIC was 527.48, whereas in the one-factor second-order model, the AIC was much higher, 726.95. Therefore, we can

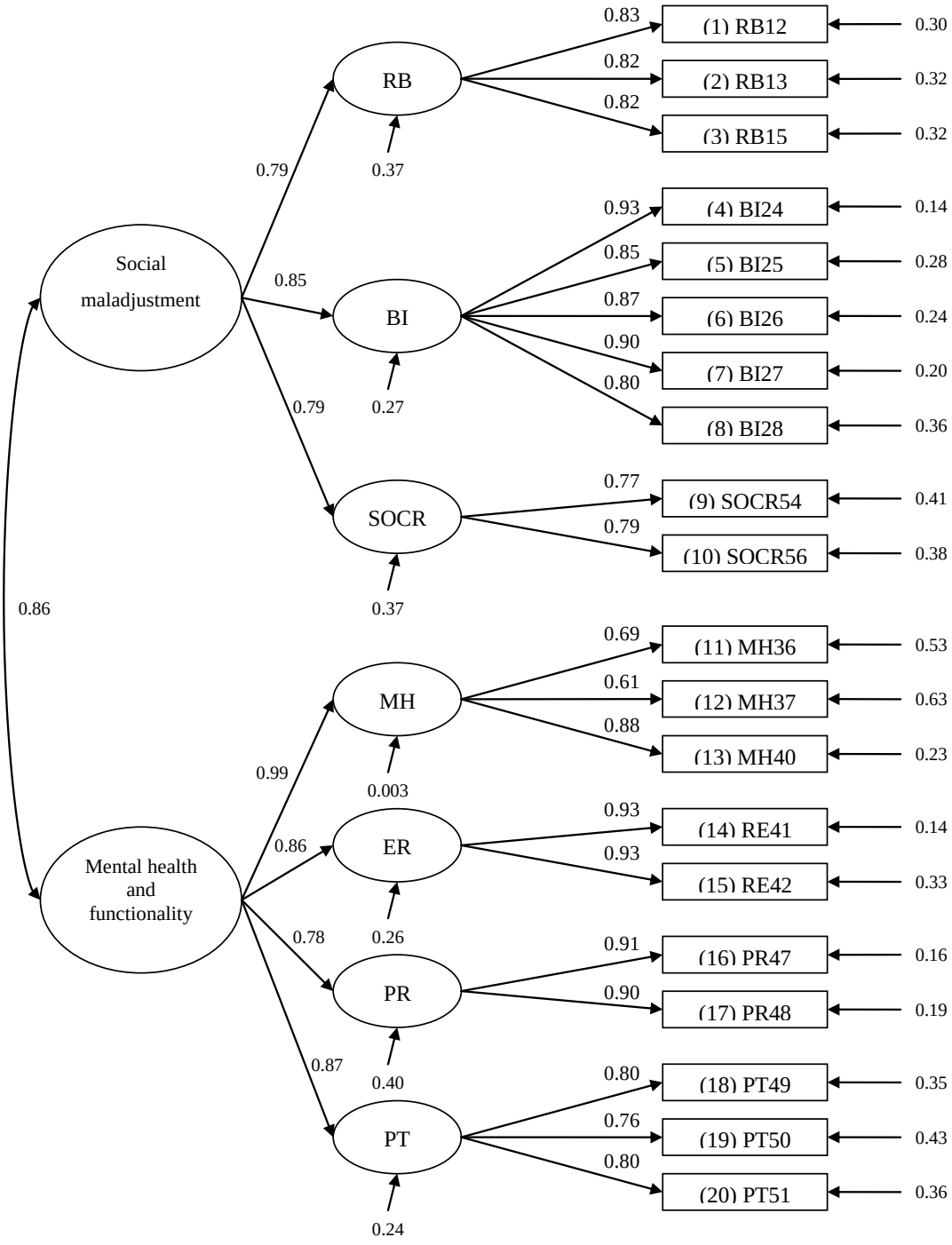


Figure 7.1. Confirmatory factor analysis of the short form, HeRQoLED-S (n = 377). RB: Restrictive behaviour; BI: Body image; SOCR: Social relations; MH: Mental health; ER: Emotional role; PR: Physical role; PT: Personality traits. Standardized parameters are shown. Results of the fit indices are as follows: $\chi^2 = 398.48$, degrees of freedom = 162, $p < 0.0001$; RMSEA (90% CI) = 0.066 (0.058, 0.074); CFI = 0.965; TLI = 0.959.

conclude that the two-factor second-order structure fits better that the one-factor second-order structure.

Then, after confirming the existence of two second-order latent traits, the GRM was applied to each factor. In a GRM, the two common item parameters are the item difficulty (β), also called the threshold, which indicates the point on the scale of the latent trait where a person has 0.5 probability of responding positively to an item category, and the item slope parameter (α), which captures the ability of an item to discriminate between people with different levels of the latent trait (Gomez, 2011). Therefore, item responses are conceptualized in terms of the slope parameter (α), and a series of $k - 1$ category thresholds, where k is the number of item response options. In our case, as the items have six response options, there will be five β parameters for the following dichotomous responses: comparing the first response option with the others, comparing the first two response options with the rest, and so on, up to comparing the first five response options with the sixth one. Therefore, each item is defined by α, which is comparable to the factor loading of classical CFA, and five β thresholds, which are indicative of the spacing of item responses along the trait dimension (Lewis and Lambert, 2006). These parameters can be used to represent the category response curve (CRC) for each item, which represents the probability of positive response to each item's response option as a function of the latent trait (Gomez, 2008).

A requirement of the IRT is that the scales must be unidimensional. The unidimensionality was tested by exploratory factor analysis, where a ratio of at least 3:1 of the eigenvalues of the first and second unrotated components was taken as evidence of unidimensionality (Gomez, 2008; Lewis and Lambert, 2006). Regarding the results of unidimensionality on each of the two dimensions, the ratio of the eigenvalues of the first and second factors obtained from exploratory factor analysis was 4.22:1 for the SocM domain and 4.78:1 for the MHF domain, both exceeding the benchmark of 3:1, indicating that the assumption of unidimensionality was not violated.

After confirming the unidimensionality, the GRM was applied to each domain separately. Given the limited sample size and low response proportions for certain categories, we chose to use a Bayesian estimator (Asparouhov and Muthén, 2010) in the GRM models. Prior assumption dependence (Lee et al., 2010) was checked by comparing slopes and category parameter estimates using different priors. This check showed that estimates were similar at a 95% confidence level regardless of the prior distribution of choice. Following the recommendations of Gelman et al. (2008) for logit models, we estimated the latent traits using weakly informative prior distributions $N(0, 1)$. Previous analysis with frequentist models yielded very high values for extreme categories in certain items, so we chose noninformative $N(0, \text{infinity})$ priors for category parameters.

For each GRM, the slope parameters (α) and thresholds (β), as well as standard errors (*SE*), were determined (Morrissey et al., 2010). In addition, the item information curve (IIC) and the test information curve (TIC) were generated (Lewis and Lambert, 2006). The fit of the GRM model was assessed as follows (Gomez, 2008, 2011; Lewis and Lambert, 2006): (a) we examined the size of each standard error estimated for each item parameter estimated; (b) the residuals of the model were determined, by comparing the observed and expected proportions of each response category of each item; and (c) for each item, a likelihood ratio χ^2 was calculated, to assess the goodness-of-fit between the expected and the observed frequencies.

Table 7.2 shows the slope (α) and threshold (β) parameters for the SocM and MHF items. Regarding slope parameters, higher values indicate greater discrimination ability of the item. The range of the latent trait that the estimated β parameters encompass indicates the adequacy of an item to represent low, medium, or high levels of a trait. In relation to the results of SocM items, α values were all acceptable and large, with *SE* less than 0.20 in all cases (which is considered acceptable) except for Item 4 (feel fat), which was slightly higher (*SE* = 0.23). The highest α values were obtained for Items 4 (feel fat), 7 (worry gaining weight), 6 (worry weight), and 5 (parts body too big), with slope parameters around 2,

Table 7.2. IRT parameter estimates from GRM for social maladjustment and mental health and functionality domains (n=377).

Item	Content	Item Parameter Estimates					
		α_1	β_1	β_2	β_3	β_4	β_5
Social maladjustment							
(1) RB12	Fast for a day, although hungry	0.85 (0.10)	0.50 (0.10)	0.87 (0.10)	1.38 (0.11)	2.13 (0.16)	2.80 (0.21)
(2) RB13	Skip some meals, although hungry	0.89 (0.08)	−0.27 (0.09)	0.11 (0.09)	0.68 (0.09)	1.39 (0.10)	1.86 (0.11)
(3) RB15	Avoid eating with others	0.75 (0.08)	−0.16 (0.09)	0.19 (0.09)	0.64 (0.09)	1.15 (0.10)	1.63 (0.11)
(4) BI24	Feel fat despite people telling you otherwise	2.26 (0.23)	−2.54 (0.25)	−1.72 (0.21)	−0.54 (0.18)	0.38 (0.19)	1.21 (0.21)
(5) BI25	Feel some parts of your body are too large	1.78 (0.16)	−2.26 (0.18)	−1.57 (0.16)	−0.67 (0.14)	0.02 (0.14)	0.76 (0.15)
(6) BI26	Worry about your weight	1.87 (0.16)	−3.50 (0.25)	−2.21 (0.18)	−1.20 (0.16)	−0.45 (0.15)	0.40 (0.14)
(7) BI27	Worry about gaining weight	2.16 (0.20)	−3.84 (0.29)	−2.87 (0.25)	−1.54 (0.19)	−0.69 (0.16)	0.21 (0.16)
(8) BI28	Avoid others seeing your body	1.02 (0.09)	−1.01 (0.11)	−0.45 (0.10)	0.23 (0.10)	0.66 (0.09)	1.25 (0.11)
(9) SOCR54	Eat negatively affecting family relations	0.54 (0.07)	−1.08 (0.09)	−0.46 (0.08)	0.17 (0.08)	0.77 (0.08)	1.76 (0.12)
(10) SOCR56	Eat negatively affecting personal relations	0.56 (0.07)	−0.88 (0.09)	−0.31 (0.08)	0.07 (0.08)	0.75 (0.08)	—
Mental health and functionality							
(11) MH36	Feel happy	1.01 (0.09)	−2.72 (0.19)	−1.68 (0.13)	−0.66 (0.10)	0.73 (0.11)	1.86 (0.12)
(12) MH37	Have mood changes that are hard to control	0.72 (0.08)	−1.65 (0.11)	−0.85 (0.09)	0.11 (0.08)	1.14 (0.09)	1.90 (0.13)
(13) MH40	Feel worthless	1.39 (0.12)	−2.00 (0.15)	−1.08 (0.11)	0.04 (0.09)	1.06 (0.12)	2.18 (0.16)
(14) ER41	Make extra effort due to emotional problems	1.84 (0.16)	−1.94 (0.16)	−0.61 (0.13)	0.92 (0.13)	1.96 (0.17)	2.88 (0.21)
(15) ER42	Accomplish less due to emotional problems	1.69 (0.13)	−1.76 (0.13)	−0.64 (0.12)	0.62 (0.11)	1.53 (0.14)	2.55 (0.17)
(16) PR47	Have difficulties paying attention due to physical problems	1.18 (0.11)	−0.87 (0.10)	0.18 (0.09)	1.21 (0.11)	2.04 (0.14)	2.69 (0.19)
(17) PR48	Stop performing tasks due to physical problems	1.07 (0.10)	−0.45 (0.10)	0.54 (0.10)	1.36 (0.12)	1.99 (0.14)	2.55 (0.18)
(18) PT49	Feel a lack of confidence in your own capabilities	1.08 (0.09)	−2.26 (0.14)	−1.15 (0.11)	−0.02 (0.09)	0.86 (0.09)	1.72 (0.12)
(19) PT50	Want to do things perfectly or not do them	0.77 (0.08)	−1.48 (0.11)	−0.83 (0.09)	−0.11 (0.09)	0.53 (0.09)	1.41 (0.10)
(20) PT51	Set very high goals and feel dissatisfied if they are not meet	0.93 (0.09)	−1.88 (0.13)	−0.79 (0.10)	0.09 (0.09)	0.62 (0.09)	1.34 (0.10)

α_1 = slope parameter estimate; β_1, β_2, β_3, β_4 and β_5 = threshold parameters. Standard error estimates for each parameter estimate are listed in parentheses.

Every question has a response scale of 6 ordinal options, from “never” to “always”.

— = This item response scale contains 5 response options.

followed by Items 1 (fast for a day), 2 (skip meals), 3 (avoid eating with others), and 8 (avoid body seen) with slope parameters somewhat lower, around 1, but still high. Finally, Items 9 (affect family relations) and 10 (affect personal relations) had much lower discriminatory ability (α = 0.54 and α = 0.56, respectively). Considering the β parameters, we can see that Items 4 (feel fat), 5 (parts body too big), 6 (worry weight), and 7 (worry gaining weight) had values of around −4 *SD* to −2 *SD* from the mean in β_1, and from around 0.25 *SD* to 1.25 *SD* from the mean in β_5, indicating that these items are particularly apparent at lower levels of severity of EDs. At the other extreme, Item 1 (fast for a day) has much higher values across all β parameters, indicating that this item only becomes apparent at higher levels of the trait. Although only shown for Item 7 (worry gaining weight) (Figure 7.2), generally in the CRCs for all items there was a noticeable shift to higher trait levels as the level of the response increased. Figure 7.3 shows the IICs for each of the SocM items and the TIC of the scale as a whole. These curves identify the position on a given trait spectrum at which the item or the scale as a whole, respectively, provides the most information. In these graphs, the *x*-axis represents the latent trait, with a scale standardized to have a mean of 0 and *SD* of 1. We can see that IIC values of Items 4 (feel fat), 5 (parts body too big), 6 (worry weight), and 7 (worry gaining weight) were relatively high across the trait spectrum compared with the rest, and in particular compared with Items 9 (affect family relations) and 10 (affect personal relations). It can also be observed that in the TIC for the scale as a whole, the highest level of information was found from −1.5 *SD* to 1 *SD* from the mean.

In relation to the GRM results for the MHF domain, Table 7.2 shows the slope (α) and threshold (β) parameters for each item. All α values were large, higher than 0.70, and with *SE* less than 0.20. The highest α values were obtained for Items 14 (extra effort) and 15 (accomplished less), with slope parameters around 1.75. In contrast, Items 12 (mood changes) and 19 (perfectionism), although providing an acceptable contribution, were less informative with slope parameters around 0.75. With respect to β parameters, most of them had a similar range of values. The only exceptions were for Items 16 (maintain attention) and 17 (stop tasks), with

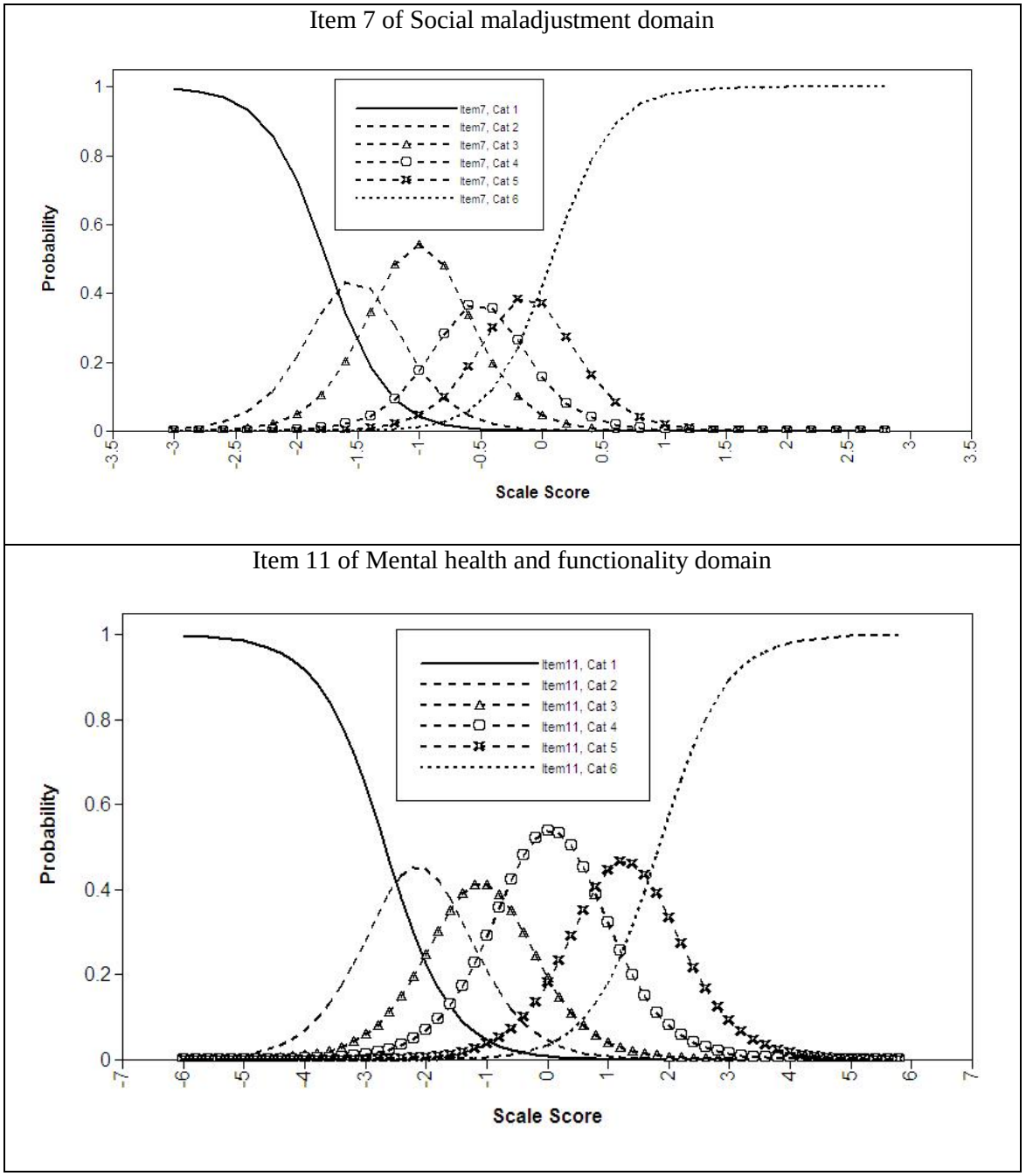


Figure 7.2. Category response curves for "Social maladjustment" item 7 and "Mental health and functionality" item 11.

slightly higher values indicating that these factors become especially apparent at higher levels of the trait. The CRCs within each of the MHF items (see Figure 7.2 for an illustration of this for Item 11 [feel happy]) showed a noticeable shift to higher trait levels as the level of the response increased. Figure 7.4 shows graphs

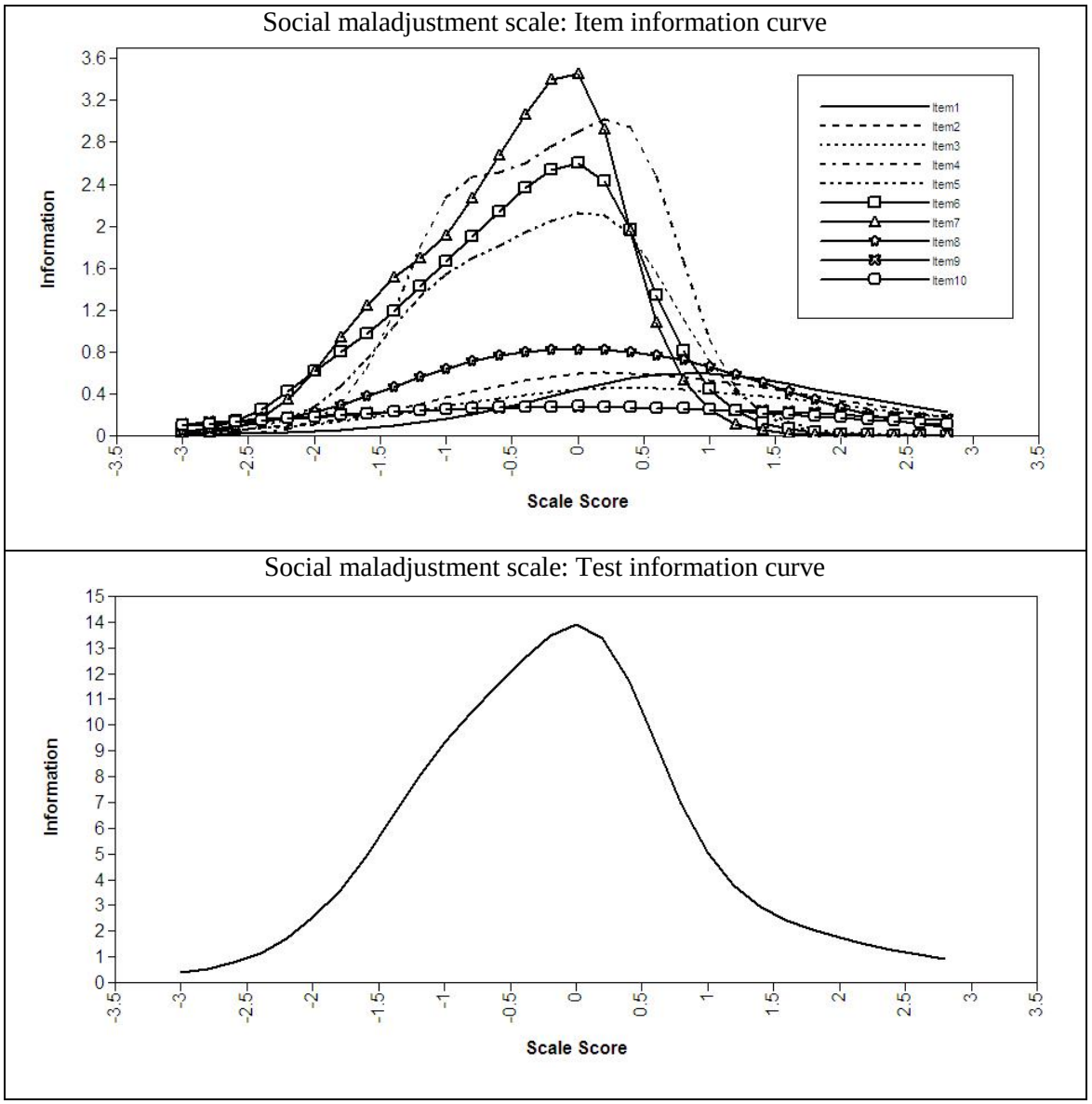


Figure 7.3. Item information curve for "Social maladjustment" items and test information curve for the "Social maladjustment" scale as a whole. As the test information curve (TIC) is the sum of the item information curves (IIC), the range of the information scale for the TIC is much larger than for the IIC.

illustrating the IICs for each MHF item and the TIC of the scale as a whole. The IIC values for Items 14 (extra effort) and 15 (accomplished less) were extremely high across the trait compared with the others, specifically from −1 *SD* to 1 *SD* from the mean. Items 12 (mood changes) and 19 (perfectionism) had the lowest IIC values across the trait spectrum, although with acceptable values. Again, as we can see in

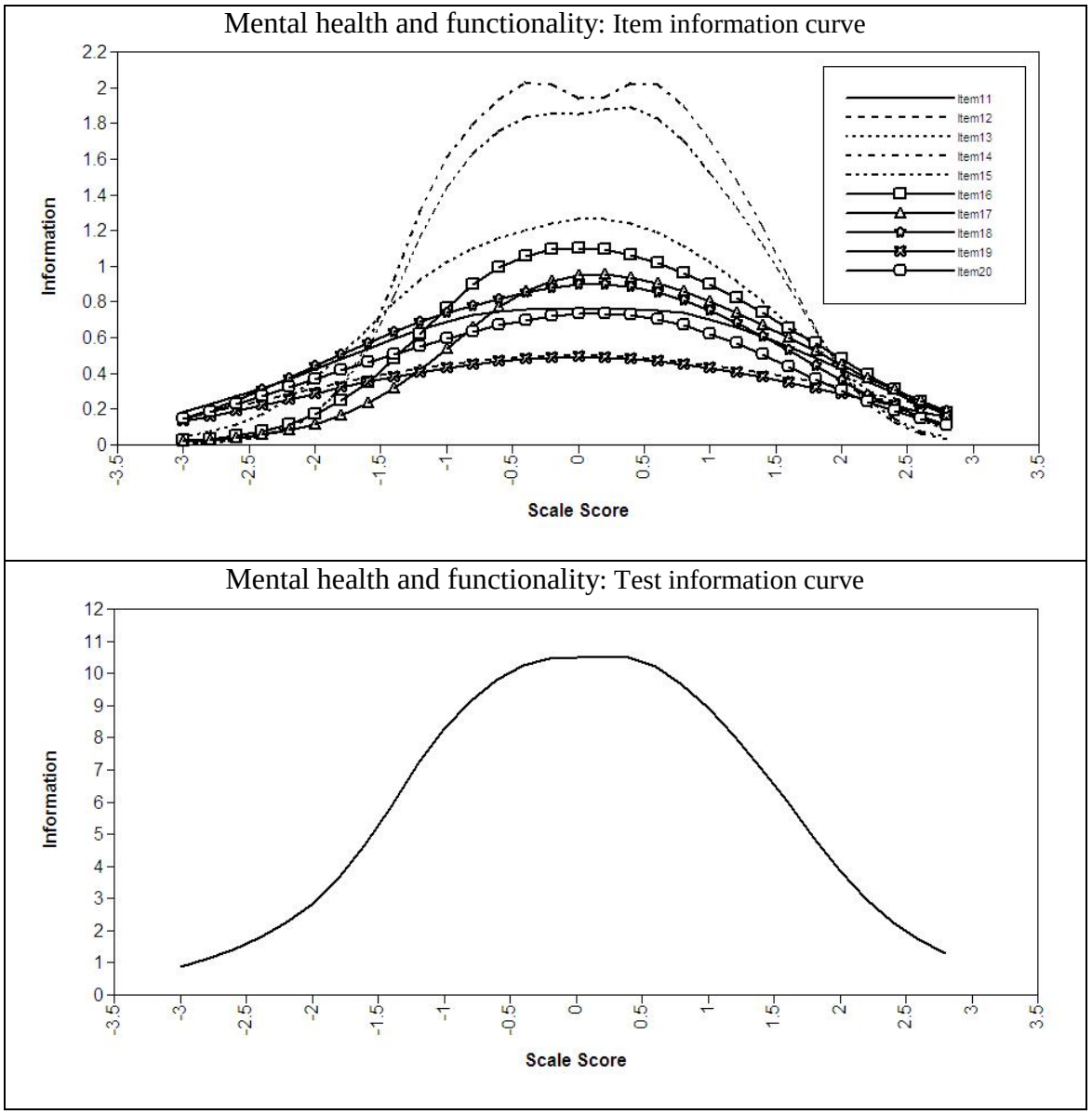


Figure 7.4. Item information curve for "Mental health and functionality" items and test information curve for the "Mental health and functionality" scale as a whole. As the test information curve (TIC) is the sum of the item information curves (IIC), the range of the information scale for the TIC is much larger than for the IIC.

the TIC for the scale as a whole, the highest level of information was provided from −1.5 *SD* to 1.5 *SD* from the mean.

Regarding the model fit, in both SocM and MHF domains, the residuals for all categories of all items were nonsignificantly different from 0, indicating a good fit.

Furthermore, there were no significant differences between the observed and expected likelihoods in any of the items of either of the domains, also showing the good fit of the models.

Convergent and discriminant validity

We hypothesized that the SocM scale would correlate positively and more strongly with the dieting concern factor of the EAT-26 than with the other domains. In the case of the MHF score, we hypothesized that it would correlate negatively and more strongly with the MCS domain of the SF-12 than with the other domains. Furthermore, the significance of a difference between two correlations was determined by the standard error of the correlation matrix ($1/\sqrt{n}$). The recommended significance criterion of two standard errors was used (Fayers and Machin, 2007). As hypothesized, the highest correlation coefficients of the SocM and MHF scales were found with the dieting concern factor of the EAT-26 and the MCS domain of the SF-12, respectively (0.81 and −0.76), being the differences between the hypothesized correlations and correlations with the other domains statistically significant (Table 7.3).

Known-groups validity

Table 7.3 shows the results of known-groups validity, in which the HeRQoLED-S mean scores were compared according to different groups of patients: (a) the CGI, where patients were classified as “mild” if he or she received a score of 0 on the CGI, “moderate” if the patient had a score of 1 or 2 on the CGI, and “severe” with a score of 3 or 4; (b) the total EAT-26 score, where patients who scored below 20 were considered to be nonsymptomatic, and those who scored 20 or above were considered to have symptoms of an ED; and (c) the MCS domain of the SF-12, where patients who scored below 50 were considered as patients with negative perception of health, and those who scored 50 or above were considered as patients with positive perception of health. We hypothesized that the more severe the ED, the higher the score on the EAT-26, and the lower the score on the MCS

Table 7.3. HeRQoLED-S correlations with the SF-12 and EAT-26 domains, and known-groups validity by comparing HeRQoLED-S mean scores according to severity groups, EAT-26 and the MCS domain of the SF-12 (n = 377).

	HeRQoLED-S	
	SocM	MHF
	r	*r*
SF-12		
MCS	−0.55	**−0.76**
PCS	−0.31	−0.43
EAT-26		
Dieting	**0.81**	0.60
Bulimia and food preoccupation	0.72	0.53
Oral control	0.40	0.33
	SocM	MHF
	$\bar{x}$ (SD)	$\bar{x}$ (SD)
Severity Index		
Minor[a]	39.12 (23.94)[b,c]	37.58 (18.30)[b,c]
Moderate[b]	49.19 (22.90)[a,c]	46.59 (19.92)[a,c]
Severe[c]	57.44 (22.79)[a,b]	53.60 (20.86)[a,b]
p value	<0.0001	<0.0001
Total EAT-26 score		
<20	26.08 (17.40)	31.53 (18.45)
≥20	62.03 (17.27)	53.82 (18.28)
p value	<0.0001	<0.0001
ES	1.45	1.04
MCS of the SF-12		
<50	52.19 (22.43)	48.70 (17.81)
≥50	26.28 (19.74)	22.28 (11.64)
p value	<0.0001	<0.0001
ES	1.05	1.24

Data are expressed as the Pearson correlation coefficient when studying the correlation between the HeRQoLED-S scales and the SF-12 and EAT-26 domains, and as the mean (SD) when comparing the HeRQoLED-S scores according to severity index, total EAT-26 and MCS of the SF-12 score groups.

SD: Standard deviation; ES: Effect size; HeRQoLED-S: Health-Related Quality of Life in Eating Disorders short form; SocM: Social maladjustment; MHF: Mental health and functionality; SF-12: 12-item Short Form Health Survey; MCS: Mental component summary; PCS: Physical component summary; EAT-26: Eating Attitudes Test-26.

[abc] Superscript letters indicate statistically differences among the three subgroups of severity index by Scheffe's test for multiple comparisons at $p < 0.05$.

domain of the SF-12, the higher patient HeRQoLED-S scores would be. For the comparison, analysis of variance with Scheffe's test for multiple comparisons or the *t* test was used, and if condition of normality was violated, the nonparametric Kruskal–Wallis test or Wilcoxon test was used. Furthermore, the magnitude of group differences was estimated by means ESs, which was estimated by calculating the mean difference divided by the pooled *SD*. Cohen's benchmarks were used to classify the magnitude of ESs: below 0.20 was not significant, between 0.20 and 0.50 small, between 0.50 and 0.80 moderate, and above 0.80 large (Cohen, 1992). We found significant differences in scores on the HeRQoLED-S scales between the three severity groups according to CGI and between the two groups according to both the total EAT-26 score and MCS of the SF-12 domain (Table 7.3). Patients with a higher level of severity, with total EAT-26 score ≥20, or with MCS score <50 had significantly ($p < 0.0001$) higher scores on both of the HeRQoLED-S scales. All ESs were above 0.80 indicating large group differences.

Reliability

Cronbach's alpha coefficient was 0.90 for the SocM scale and 0.91 for the MHF scale, that is, in both cases it was higher than the minimum value of 0.70, which is considered acceptable (Nunnally and Bernstein, 1994).

Responsiveness

First, we compared principal characteristics between patients who responded to the follow-up and those who did not. Chi-square or Fisher's exact tests were performed to compare categorical variables, while *t* tests were used to compare continuous variables, or the Wilcoxon nonparametric tests if normality was violated. Of the 132 patients in the second cohort, 85 responded to the follow-up after 1 year (64.39%). The only significant differences between the participants who responded to the follow-up and those who did not were found in dieting concern and oral control factors of EAT-26, with patients who did not respond to the follow-up having lower ED symptomatology at baseline.

To study responsiveness patients were classified as "improved" or "worsened" based on the change scores of the total EAT-26 score as follows: patients with a change score greater than zero were classified as "improved," and those with a change score of less or equal to zero were classified as "worsened." Then, ceiling and floor effects at baseline and at 1 year were examined to evaluate the discriminatory ability of the scales. Means for the HeRQoLED-S scales in each group of patients at baseline and after 1 year were compared by means of a paired *t* test or nonparametric Wilcoxon signed-rank test. Finally, to measure the responsiveness of the HeRQoLED-S, we used the SES, defined as the mean change score divided by the *SD* of the baseline scores, and SRM, defined as the mean change score divided by the *SD* of the change scores. Cohen's guidelines were used to interpret the magnitude of the ESs (Cohen, 1992). Table 7.4 shows the results of responsiveness analyses. Both the HeRQoLED-S scales showed minor floor and ceiling effects (<5%) at baseline and after 1 year of follow-up. Among patients classified as "improved," the SocM and MHF scales increased by 12.01 and 9.26 points after 1 year of follow-up, respectively, both of which were significant changes ($p < 0.001$). On the other hand, among patients classified as "worsened," the change scores after 1 year decreased −5.25 and −4 points, respectively. The SES and SRM responsiveness parameters among patients classified as "improved" were between 0.57 and 0.61 for SocM scale and between 0.43 and 0.52 for MHF scale, indicating moderate changes. Responsiveness parameters among patients classified as "worsened" were, however, much lower, from −0.47 to −0.18, indicating nonsignificant to small changes.

7.5. Discussion

The results of the current prospective study with a new cohort of patients and a considerable sample size provide more evidence of the reliability and validity of the HeRQoLED-S, and the results of patients from Cohort 2 followed-up after 1 year provide a first approach on its responsiveness. To the best of our knowledge, this is the first ED-specific questionnaire submitted to a rigorous psychometric analysis, using a new relatively large cohort of patients.

Table 7.4. Responsiveness parameters one year after in the HeRQoLED-S domains in cohort 2.

	HeRQoLED-S			
	Improved (n = 63)		Worsened (n = 21)	
	SocM	MHF	SocM	MHF
% at floor				
Baseline	0	0	0	0
Follow-up	0	0	0	0
% at ceiling				
Baseline	1.61	0	0	0
Follow-up	1.59	0	0	4.76
$\bar{x}$ (SD)				
Baseline	58.69 (21.20)	51.32 (21.59)	47.91 (24.63)	44.38 (22.84)
Follow-up	46.61 (25.27)	42.03 (22.16)	53.16 (23.08)	48.38 (24.10)
Change	12.01 (19.56)†	9.26 (17.90)‖	−5.25 (11)†	−4 (14.11)‖
p value*	<0.0001	0.0001	0.0332	0.4159
SES	0.57	0.43	−0.21	−0.18
SRM	0.61	0.52	−0.47	−0.28

* Paired *t*-test or nonparametric Wilcoxon signed rank test to compare the baseline and follow-up mean scores.

† Statistically significant differences in "social maladjustment" change scores between patients classified as "improved" and those classified as "worsened" ($p = 0.0003$).

‖ Statistically significant differences in "mental health and functionality" change scores between patients classified as "improved" and those classified as "worsened" ($p = 0.0014$).

HeRQoLED-S: Health-Related Quality of Life in Eating Disorders short form; SocM: Social maladjustment; MHF: Mental health and functionality; % at floor: Percentage of the study population at the lowest possible scale level; % at ceiling: Percentage of the study population at the highest possible scale level; SD: Standard deviation; SES: Standardized effect size; SRM: Standardized response mean.

Changes were calculated by subtracting follow-up scores from baseline score, with a positive result indicating a gain.

It is very useful to have short questionnaires that are reliable, valid, and responsive, in order to collect information correctly while placing less burden on patients and physicians. With the findings of the present study, we provide further evidence that the HeRQoLED-S has these characteristics, ensuring that the instrument can be used in epidemiological studies and daily clinical practice, and suggesting that it could be a very good alternative to the long version.

As stated previously, the first ED-specific questionnaires related to HRQoL were developed almost simultaneously (Abraham et al., 2006; Adair et al., 2007; Engel et al., 2006; Las Hayas et al., 2006). However, to the best of our knowledge, ours is the only one that has been reduced and validated in a new sample of ED patients different from the sample used for its derivation, thus providing further evidence of its reliability and validity. In addition, the new sample of 377 ED patients we used is a large sample given the low incidence of this disease, and this fact gives robustness to the results.

In general, to assess the progress of an ED, it is more common to use symptom scales and scales that measure behavioral changes than HRQoL measures. However, by the very nature of the ED, it has a very negative impact on the typical areas of HRQoL, such as physical, mental, and social domains. Hence, it is important to have available measures of quality of life in ED. The use in clinical practice of the HeRQoLED-S may be very varied, from the general assessment of the state of the person on the first day of consultation to its use throughout the treatment to assess its progress (Abraham et al., 2006; Adair et al., 2007; Engel et al., 2006; Hay and Mond, 2005; Las Hayas et al., 2006; Padierna et al., 2002). Moreover, in case of employing different treatments, an accurate measure of HRQoL can provide information on which of them produces a greater improvement in the HRQoL. Hence, it is important to create measures that are both valid with respect to the construct being measured and sensitive to change. In this new study, the construct validity has been studied using more appropriate statistical techniques given the ordered-response categories of the items. Taking into account all the results of the CFA and GRM for construct validity, the overall findings here suggest that on the whole the HeRQoLED-S has good psychometric properties from both classical and IRT perspectives and therefore would be useful for measuring HRQoL in ED patients.

Specifically, the assessment of the HeRQoLED-S instrument using CFA specific for categorical variables strengthens the evidence of its construct validity. In the derivation study of the HeRQoLED-S (Las Hayas et al., 2010), we used classical

CFA, which requires continuous measures and multivariate normality. Though six response options tend to produce very similar results in CFA to continuous responses, using CFA for categorical variables provides greater robustness and reliability to our results in this new study. Notably, the results obtained from the CFA fit indices were similar, or even slightly higher, than those obtained in the derivation study (Las Hayas et al., 2010).

Regarding the IRT analysis, the application of the GRM to the HeRQoLED-S has aimed particularly to improve the psychometric qualities. We used the GRM procedure, which again is recommended for the evaluation of polytomous items (Maydeu-Olivares, 2005), and consequently the results obtained are more robust, and its application has served to confirm that the questionnaire has adequate internal structure that gives more validity to the results. As a general result of GRM, we can say that all the items were good at discriminating different levels of the relevant latent traits (SocM and MHF) and had reasonable information values (reliability) for these traits from the mean level and fairly symmetrical distributions around the mean. Another benefit of the application of IRT to the HeRQoLED-S is that we now can assure that the questionnaire items work with stability, so the measures will not vary from sample to sample (Hays et al., 2000). However, there were some specific findings worth noting. Items 9 (affect family relations) and 10 (affect personal relations) among the SocM domain and Items 12 (mood changes) and 19 (perfectionism) of the MHF domain had relatively lower slope parameter values, which is consistent with the results of the CFA, given that (although they were high) the lowest factor loadings among each domain were found in these items. Instead, items related to obsession with weight and body size, such as Item 4 (feel fat), Item 7 (worry gaining weight), Item 6 (worry weight), and Item 5 (parts body too big), contributed the most to the definition of the SocM scale. Therefore, the name of the subscale "Social Maladjustment" may be improved to be more appropriate to the items that contribute most to the scale. Future studies may explore this issue further, and change the name to "Obsessive traits." Regarding the MHF domain, the items that contributed most to the

construct were Item 14 (extra effort) and Item 15 (accomplish less), which reflect the functional limitations arising from suffering a serious emotional problem.

In relation to the threshold values of the SocM domain, Item 1 (fast for a day) had the highest values, indicating that this item become more apparent at higher levels of the trait. This finding is consistent with the results obtained from the Rasch analysis in the previous study of the derivation of HeRQoLED-S (Las Hayas et al., 2010), in which this item was plotted in the highest positions of the corresponding item-person maps. Studies by Mond et al. (2005) and Padierna et al. (2000) showed that patients diagnosed with anorexia nervosa restrictive subtype presented the highest levels of quality of life, in comparison with other ED subtypes. Mond et al. (2005) justifies their results arguing that anorexia nervosa patients find being thin as rewarding socially and internally so they do not perceive subjectively that fasting for a day may decrease their quality of life. Nevertheless, the HeRQoLED-S, after GRM analysis, identifies the item about fasting as the most severe scenario, which may have a very negative impact on HRQoL of the individual. This result is in line with studies that assess the serious consequences of starvation on health of ED individuals. de la Rie et al. (2005) noted the consequences of starvation on the health of people with EDs who practice dietary restraint. Specifically, they state that people with anorexia nervosa may have low bone density, fertility problems, and heart and kidney abnormalities. Fasting also has an effect on cognitive and behavioral level toward promoting behaviors such as binge eating and psychological manifestations such as increased emotional responsiveness, dysphoria, and distractibility (Polivy, 1996).

Regarding the MHF domain, items that become visible in negative states of quality of life and indicate greater severity are those that refer to suffer functional limitations due to poor physical health, such as Items 16 (maintain attention) and 17 (stop tasks). As stated by de la Rie et al. (2005), physical deterioration caused by the disorder can be very limiting and disabling. Physical limitations associated with binge eating and vomiting behaviors typical of bulimia nervosa are gastric problems, esophagus, and often dental erosion (de la Rie et al., 2005). The more

frequent the binge eating and vomiting, the greater the impact on the physical health of the person. A fragile physical health impacts negatively in all areas of life quality, as shown by severity thresholds detected in this domain.

Future studies could further explore and improve the psychometric qualities of HeRQoLED-S by developing new items that can cover the gaps in each of the factors of HRQoL that are not measured by any of the scale items. It would also be interesting to explore whether the application of GRM improved measurement accuracy in comparison with the classical test theory. This would be possible by calculating the HeRQoLED-S scores as classical test theory and IRT in the same sample and estimating the relative precision index, which indicates how much more or less precise a new scoring method (in this case, the IRT-based score) is relative to the standard (in this case, the summative-based score) in distinguishing groups expected to differ.

Analysis of the internal consistency shows acceptable levels of reliability, and the convergent validity of the HeRQoLED-S was demonstrated by confirming the hypothesis of correlations with the EAT-26 and SF-12 domains, and the HeRQoLED-S had an excellent known-groups validity. These results were practically the same as those obtained in the derivation study (Las Hayas et al., 2010), indicating that the short form maintained its good internal consistency and validity.

The HeRQoLED-S showed acceptable responsiveness at 1 year of follow-up. Compared with the results obtained in the responsiveness study of the long version of the HeRQoLED (Las Hayas et al., 2007), the responsiveness parameters for SocM domain improved from small changes in the long form to moderate ones in the current short form. For the MHF domain we also found slightly higher values for the responsiveness parameters than those obtained for the corresponding first-order factors of the long questionnaire (Las Hayas et al., 2007). Among patients classified as “worsened” the results are less informative given that the sample size is small (n = 21). Even so, the results for the SocM domain were slightly higher

than those obtained for the corresponding first-order factors of the long form of the questionnaire. We recognize that, as was the case with the long version, the evidence for the responsiveness is not very strong, but as Strober et al. (1997) underlined recovery time in patients with ED is long, between 57 and 79 months, depending on the definition of recovery. Padierna et al. (2006) also reported that after 2 years of follow-up only 10% of the patients recovered completely or partially. In line with this, we attribute the moderate responsiveness results obtained in our study to the follow-up being only 1 year, and that this may not be enough time for even partial recovery.

A further limitation is the lack of a test–retest study to provide a complete assessment of its reliability. Moreover, taking into account that the two latent factors are highly correlated, and given the modest results obtained for discriminant validity, future work should be performed to assess the discriminant validity of the SocM and MHF scales. Regarding responsiveness, missing data are a key limitation of the prospective cohort design and a common finding when conducting follow-up studies (Abraham et al., 2006; Bohn et al., 2008; Las Hayas et al., 2007). In our case, there was a quite good response rate at 1 year of follow-up (almost 65%). The losses occurred despite our mailing up to two reminders and contacting nonresponders by telephone. Furthermore, no differences were observed in the key variables, namely, SocM and MHF baseline scores, when responders were compared with nonresponders. Therefore, although a bias may have been present in our responsiveness study due to missing data, it is likely to be small and we believe the results can be generalized to the entire sample. Another limitation in our responsiveness study is that only one of the two cohorts was followed-up, and consequently, the sample size was greatly reduced for analysis. Future work should, therefore, include extending the study of responsiveness with a larger cohort of ED patients and also consider the inclusion of a transition question for each domain in the follow-up questionnaire. This latter addition would make it possible to assess the responsiveness according to whether patients had improved, their status had not changed, or they had worsened with respect to their baseline HRQoL, in terms of the SocM and MHF domains. Another limitation

could be the use of the EAT-26, given that despite being a widely used tool, it has also been criticized for its unstable internal structure and scoring system (Maïano et al., 2013) and, finally, the lack of a predictive validity study of the HeRQoLED-S.

In conclusion, this comprehensive validation process, which used a new and sufficiently large cohort of ED patients, combining classical and more modern methods, such as GRM models, showed that the HeRQoLED-S is a valid and reliable tool for measuring SocM and MHF domains in patients with ED and provides evidence of its moderate responsiveness at 1 year of follow-up. Its simplicity and ease of application will increase its acceptability and usefulness within the psychiatry and psychology community, and in particular, it may be of interest in routine practice given that a goal in that context is to collect information involving as little effort as possible for both the patient and the physician. In clinical research, where patients usually have to complete several questionnaires implying a great burden, short questionnaires result in improved patient compliance and response rates, and therefore, this shorter version will further enhance its applicability. In conclusion, the HeRQoLED-S is a good alternative to the long form maintaining good psychometric properties.

Capítulo 8

Discusión general

En la presente tesis se utilizan diferentes métodos estadísticos, tanto de la teoría clásica del test (TCT) como de la teoría de la respuesta al ítem (TRI), para validar y/o reducir tres cuestionarios específicos de calidad de vida relacionada con la salud (CVRS) para tres ámbitos distintos de la salud: patología endocrinológica, osteomuscular, y psiquiátrica. Concretamente, se ha utilizado el cuestionario Obesity-related Problems scale (OP) diseñado para medir el impacto de la obesidad mórbida en el funcionamiento psicosocial; el cuestionario reducido del Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), empleado para la medición de sintomatología y función en pacientes con artrosis de cadera; y el cuestionario Health Related Quality of Life for Eating Disorders version 2 (HeRQoLEDv2) diseñado para medir la CVRS en pacientes con un trastorno de la conducta de la alimentación. Los resultados de este trabajo demuestran por un lado, que los métodos estadísticos basados en la TCT y los basados en la teoría moderna o TRI son complementarios entre sí. Cada uno de

ellos proporciona diferente tipo de soporte en el proceso de validación y reducción de cuestionarios de CVRS. Por otro lado, se muestra que la combinación de los diferentes métodos estadísticos en el proceso de validación y/o reducción de los tres cuestionarios específicos de CVRS evaluados, proporciona evidencia de la validez de estos instrumentos y de sus buenas propiedades psicométricas, proporcionando así versiones científicamente validadas en castellano que pueden ser utilizadas en la práctica clínica, así como entre investigadores y decisores sanitarios.

En el proceso de validación y/o reducción de estos tres cuestionarios, hemos analizado diferentes propiedades psicométricas, tales como la fiabilidad, la validez y la sensibilidad al cambio. A su vez, dentro de la validez hemos analizado diferentes subtipos de validez, validez de grupos conocidos, validez convergente y discriminante, o validez de estructura, englobados dentro de la validez de constructo, tal y como se muestra en la Figura 1.1 del Capítulo 1. Como ya hemos comentado anteriormente, aunque la validez de constructo es la más susceptible de exploración por el análisis estadístico, los métodos estadísticos a utilizar para la evaluación de la validez convergente y discriminante, o la validez de grupos conocidos, o incluso la validez de criterio o la fiabilidad, son métodos bien establecidos y ampliamente aceptados. Esto no ocurre con la validez de estructura. Por tanto, en este apartado nos centraremos en la discusión de los diferentes métodos utilizados para el análisis de la validez de estructura, ya que la existencia de diferentes perspectivas y teorías de enfoque, hace que sea un tema más controvertido.

Así, en este Capítulo 8, presentaremos una discusión general, ya que en cada uno de los capítulos de resultados (Capítulos 4, 5, 6 y 7) ya se ha presentado la discusión específica de cada uno de los estudios de validación y/o reducción de los tres cuestionarios de CVRS evaluados en esta tesis doctoral. En primer lugar presentaremos la discusión general basándonos en los resultados presentados en los capítulos 4, 5, 6 y 7, discutiendo los diferentes métodos estadísticos utilizados para la evaluación de estructura, planteando las ventajas y limitaciones de cada

uno de los métodos, para así proporcionar algunas recomendaciones a la hora de validar y/o reducir cuestionarios de CVRS. En segundo lugar, discutiremos brevemente los resultados psicométricos de fiabilidad, validez y sensibilidad al cambio obtenidos para cada uno de los cuestionarios evaluados en la presente tesis, los cuestionarios OP, WOMAC, HeRQoLEDv2, y Health Related Quality of Life for Eating Disorders-Short Version (HeRQoLED-S). Y finalmente, plantearemos las líneas de investigación futura.

8.1. Métodos para la evaluación de la validez de estructura

8.1.1. Métodos de la teoría clásica del test y de la teoría de la respuesta al ítem

Métodos basados en la teoría clásica del test

Entre los métodos de la TCT hemos empleado el Análisis Factorial Exploratorio (AFE), el Análisis Factorial Confirmatorio (AFC) y los Modelos de Ecuaciones Estructurales (MEE), tanto clásicos como para datos categóricos. El AFE es un método meramente exploratorio, con las limitaciones que ello supone. Aunque el hecho de que sea únicamente exploratorio proporcione menor evidencia en la validez de estructura de un cuestionario de CVRS, es el método más común y el más utilizado dentro de los estudios de validación de cuestionarios de CVRS. Esto se debe a que es un método simple, que no requiere de ninguna especificación previa acerca de la teoría de medición del cuestionario, ni de la interrelación entre los ítems y los constructos, y que además está implementado en cualquier programa estadístico. Sin embargo, una de sus limitaciones, es la poca flexibilidad que permite al establecer las interrelaciones entre ítems y constructos, o entre los propios constructos. En relación a las interrelaciones entre ítems y constructos, el AFE proporciona pesos de todos los ítems en todos los factores (Figura 3.1). La magnitud de dicho peso es el que ayuda a decidir cuál es la importancia de cada ítem en cada factor. Por tanto, no podemos considerar ítems que solo tengan peso en unos factores y no en todos. En relación a las interrelaciones entre constructos,

el AFE únicamente dispone de dos opciones: todos los factores están incorrelacionados entre sí (rotaciones ortogonales) o todos los factores están correlacionados entre sí (rotaciones oblicuas). Por tanto, no es posible establecer relación entre algunos factores y no entre todos. Esta limitación referente a las interrelaciones es especialmente más notable cuando el cuestionario es multidimensional. En cuestionarios unidimensionales, esta limitación no es un problema. En nuestro caso hemos aplicado el AFE únicamente al cuestionario OP, siendo un cuestionario con una estructura sencilla y unidimensional, con lo que algunas de estas limitaciones como la falta de flexibilidad para ajustar estructuras multidimensionales no son tan notorias. A pesar de estas limitaciones, este método es de gran utilidad en caso de desconocimiento sobre la estructura interna subyacente de un instrumento, ya que ayuda a conocer el número de factores (o constructos) existentes, y a conocer como se relacionan los ítems con los factores. Aunque esto no suele o no debería ser algo habitual, ya que cuando se desarrolla un instrumento siempre se debe tener clara la teoría de medición, en algunas ocasiones puede ocurrir que la estructura hipotetizada no se confirma y el AFE puede ayudar a establecer relaciones. En los cuestionarios específicos, donde se pretenden medir a través de diferentes preguntas constructos relacionados con la enfermedad, muchas veces esos constructos están a su vez muy relacionados entre sí, y algunas veces puede resultar difícil diferenciarlos. Así, puede ocurrir que las estructuras hipotetizadas no se confirmen, y el AFE nos puede ayudar a entender el por qué no se están cumpliendo, proporcionándonos pistas de cómo se agrupan realmente los ítems.

Otro método que hemos utilizado es el AFC, que aunque también es un método clásico, es algo menos utilizado que el AFE. A diferencia del AFE, este sí que es un método confirmatorio, lo que resulta ser de mucha mayor utilidad en el proceso de validación de estructura de instrumentos de CVRS. El hecho de obtener buenos resultados en el AFC nos proporciona mucha mayor evidencia de la validez de estructura del cuestionario que si lo hubiéramos hecho mediante un AFE. Por otro lado, es un método también mucho más flexible que el AFE ya que en este caso, permite establecer qué ítems se ven afectados por cuales factores. No es necesario

establecer relaciones de todos los ítems con todos los factores. Un ítem se puede ver afectado por uno o más factores. Además, podemos establecer relaciones entre los factores latentes o no, ya que nosotros decidimos qué factores latentes están relacionados y cuales no (Figura 3.2). Estos métodos son adecuados tanto para cuestionarios unidimensionales como multidimensionales, aunque siempre que el cuestionario no tenga estructuras complejas del tipo estructuras de segundo orden, o variables causales (Figura 3.3), para lo que se utilizan los MEE.

Los MEE son métodos de la psicometría clásica, más sofisticados, que han sido sobre todo utilizados en el ámbito de la educación, así como en los test de personalidad, aunque no se utilizan tanto en estudios de CVRS. Estos modelos son la generalización de los modelos factoriales que permiten ajustar estructuras de cuestionarios mucho más complejas, además de permitir ajustar modelos causales. Por lo tanto, los MEE podrían proporcionar una muy buena alternativa a los métodos más clásicos como el AFE y AFC, cuando estos no son capaces de ajustar la estructura subyacente del cuestionario. Un claro ejemplo de cuestionarios con estructuras más complejas de segundo orden es el cuestionario HeRQoLEDv2. Tal y como se muestra en la Figura 1.7 del Capítulo 1, 40 de los ítems se conforman en siete factores, denominados factores de primer orden, y a su vez estos siete factores se resumen en dos medidas resumen, denominados factores de segundo orden. En estos casos es fundamental utilizar métodos como los MEE para comprobar la validez de estructura del instrumento.

Como ya hemos comentado, el AFC o los MEE son métodos confirmatorios en los que debemos de establecer a priori una estructura del cuestionario, para así mediante los índices de bondad de ajuste decidir si el modelo es o no adecuado. En el AFE nos basamos en la magnitud de los pesos factoriales de cada ítem con cada factor latente para decidir qué ítems se ven afectados por cuales factores. En los AFC y los MEE, por un lado disponemos de los índices de bondad de ajuste que nos indican si el modelo planteado es correcto o no. Pero, estos índices nos proporcionan información de manera global, con lo que si el modelo no ajusta bien, debemos hacer modificaciones, eliminar relaciones contempladas o añadir nuevas

relaciones no contempladas, hasta encontrar el modelo adecuado. Para eliminar relaciones contempladas, nos podemos basar al igual que en el AFE en los pesos factoriales obtenidos, de forma que si los pesos son bajos o no significativos, significa que esa relación establecida no es adecuada. Por otro lado, para añadir nuevas relaciones podemos basarnos en los resultados de los multiplicadores de Lagrange que consisten en estimar la reducción del estadístico χ^2 de bondad de ajuste, al añadir una nueva relación no contemplada previamente (Hatcher, 1994). Por tanto, todas estas herramientas hacen que el AFC o los MEE sean métodos muy flexibles y además permiten llegar a la solución. La limitación que estos métodos pueden tener son la dificultad de establecer las hipótesis previas sobre la estructura subyacente del cuestionario, y que son necesarios programas estadísticos más específicos para su aplicación, ya que no suelen estar implementados en los paquetes estadísticos de uso más general, lo que hace que sean métodos menos utilizados. Sin embargo, si los resultados del AFC o MEE son buenos, esto garantiza la validez de estructura del instrumento, proporcionando mucha mayor evidencia de dicha validez que lo que se obtendría mediante un AFE.

Entre estos modelos factoriales, todavía hay algunas cuestiones sin resolver, siendo una muy importante el supuesto de normalidad de los ítems. Estas técnicas se basan en el supuesto de que los ítems siguen una distribución normal. Sin embargo, la mayoría de los cuestionarios de CVRS constan de ítems con escalas de respuesta tipo Likert, es decir, variables categóricas y no normales. Los métodos clásicos de análisis factorial utilizan en general el método de estimación de máxima verosimilitud basándose en la matriz de correlaciones de Pearson, con errores distribuidos de manera normal. Si esta asunción es violada, los test de bondad de ajuste no son fiables. Sin embargo, debido a que la mayoría de la gente emplea técnicas clásicas, también hay estudios de cómo de robustas son estas técnicas ante ciertos incumplimientos de los requisitos. Hay autores que dicen que se pueden tratar las variables como si fueran continuas si tenemos al menos cinco opciones de respuesta, la muestra es suficientemente grande y los datos no son extremadamente sesgados (Cohen y cols., 2003). Para definir si son o no extremadamente sesgados, podemos utilizar los coeficientes de kurtosis y

asimetría. Un valor absoluto del coeficiente de asimetría superior a 3 indica que la distribución es extremadamente sesgada. Un valor absoluto del coeficiente de kurtosis superior a 10 indica que tenemos un problema, y si es superior a 20 indica que tenemos un serio problema (Kline, 2005). En cualquier caso, ignorar el hecho de la normalidad y asumir que las variables son continuas cuando realmente son categóricas con respuestas de tipo Likert puede dar lugar a resultados sesgados en caso de utilizar métodos de estimación como por ejemplo el de máxima verosimilitud (Raykov y Marcoulides, 2006). Ello conlleva a que un mal resultado en los índices de ajuste del AFC o los MEE puede significar que realmente el modelo planteado no es adecuado, o que las variables no cumplen los requisitos necesarios con lo que los resultados que obtenemos son de poca fiabilidad. Por tanto, aunque los métodos clásicos son robustos ante un ligero incumplimiento de normalidad, si tenemos una distribución extremadamente no normal, es mejor no utilizar este método de estimación (Harrington, 2009). Para solventar este problema se ha desarrollado metodología para la utilización del análisis factorial y los MEE para datos categóricos (Lee y cols., 1995), aunque también es cierto que esta metodología no está implementada en la mayoría de los paquetes estadísticos, con lo que su uso es aún mucho más reducido. Cuando no podemos usar el método de máxima verosimilitud como estimador, debemos usar otros estimadores para datos categóricos (Harrington, 2009), como por ejemplo un estimador robusto de máxima verosimilitud, o un estimador robusto de mínimos cuadrados ponderados, entre otros. Otros autores han planteado la utilización de métodos de distribución libre pero sin embargo para ello se requieren muestras muy grandes. Por ejemplo, un análisis para 15 variables con 3 factores requeriría muestras de entre 2500 y 5000 observaciones (Harrington, 2009). Algunos autores también han sugerido la utilización de la matriz de correlación policórica para datos categóricos. Sin embargo para estos casos también se recomienda la utilización de muestras mayores (Fayers y Machin, 2007).

Otro aspecto importante a comentar es la muestra necesaria para los análisis factoriales. Parece haber un acuerdo generalizado en que a mayor muestra mejor será para el análisis factorial. Sin embargo no hay un consenso sobre cuál es la

muestra mínima necesaria, pero es cierto que dependerá en gran parte del número de ítems que tenga el instrumento y del grado de complejidad de su estructura. Existen diferentes aproximaciones al tema del tamaño muestral necesario, desde "reglas de oro" hasta estudios basados en simulaciones de Monte Carlo. En cuanto a las reglas de oro, por ejemplo una muestra inferior a 100 sujetos se considera poca muestra e inapropiado; entre 100 y 200 sujetos se considera un tamaño muestral medio y aceptable si el modelo no es muy complejo; y más de 200 sujetos se considera una muestra grande y aceptable para la mayoría de los modelos (Kline, 2005). Otra regla de oro, y de las más utilizadas en instrumentos de CVRS, es la de un mínimo de entre 5 o 10 veces el número de ítems, aunque realmente tampoco existen bases teóricas en las que se basen estas reglas (Fayers y Machin, 2007). En estos casos siempre conviene ser conservadores, y considerar un mínimo de 10 veces el número de ítems (Arrindell y Van der Ende, 1985; Nunnally, 1978; Velicer y Fava, 1998). Por otro lado, Lee y Song (2004) llevaron a cabo un estudio de simulación comparando el método de máxima verosimilitud y el estimador bayesiano con muestras pequeñas. Concluyeron que el método de máxima verosimilitud no era recomendado con muestras pequeñas a pesar de tener datos normales. Aun así, recomiendan el uso del estimador Bayesiano siempre que la muestra sea de 2 a 3 veces el número de parámetros a estimar. También existen estudios (Lee y Song, 2004), que demuestran que los métodos basados en distribución libre para datos no normales requieren de muestras mucho mayores que los métodos con estimadores de máxima verosimilitud para datos normales. Muthén y Muthén (2002) sugieren que hay que tener cuidado con las reglas de oro referentes al tamaño de muestra necesario ya que la muestra mínima necesaria depende de muchos factores, como por ejemplo, el tamaño o complejidad del modelo, la distribución de las variables, la fiabilidad de las variables, y el grado de fuerza existente en las relaciones entre variables. Por ejemplo, Muthén y Muthén (2002) en un estudio de simulación, encontraron que para su modelo necesitaban una muestra de 150 observaciones cuando los datos se distribuían de manera normal, y que esta muestra aumentaba a 265 cuando no se distribuía de manera normal. Este dato pone de manifiesto el impacto que tiene por ejemplo la distribución en el cálculo del tamaño muestral.

Métodos basados en la teoría de la respuesta al ítem

Entre los métodos de la TRI hemos empleado los modelos para datos politómicos. Los modelos logísticos de 1, 2 o 3 parámetros suelen ser menos utilizados que su generalización para datos politómicos en el proceso de evaluación de la validez de estructura de cuestionarios de CVRS, debido a que la mayoría de estos cuestionarios están formados por ítems politómicos con respuestas categóricas tipo Likert. Sin embargo, también es cierto que en muchas ocasiones nos solemos referir a la aplicación de análisis Rasch cuando se aplica un modelo de la TRI, independientemente del modelo que sea, debido a su inventor Rasch (1960). En cuanto a las extensiones de los modelos logísticos de 1–parámetro para ítems de respuesta politómica, tenemos los modelos Partial Credit Model (Masters, 1982) y Rating Scale Model (Andrich, 1978), siendo este último un caso particular del anterior cuando se asume que el "step" de dificultad en las categorías o niveles del ítem es constante. De esta forma, el Rating Scale Model es útil cuando todos los ítems de la escala tienen el mismo número de opciones de respuesta. En caso contrario, es necesario utilizar el Partial Credit Model.

Dentro de las extensiones de los modelos logísticos de 2–parámetros para ítems de respuesta politómica tenemos el Generalized Partial Credit Model (Muraki, 1992), el Graded Response Model (Samejima, 1969), y el Nominal Response Model (Bock, 1972). Este último prácticamente no se utiliza en la evaluación de cuestionarios de CVRS ya que no es habitual encontrarnos cuestionario con ítems de respuesta nominal en este ámbito. Así, los métodos de 2–parámetros más utilizados para datos categóricos son el Generalized Partial Credit Model y el Graded Response Model. La diferencia entre ellos radica principalmente en que el Graded Response Model requiere que los ítems tengan categorías de respuesta ordenadas, mientras que en el Generalized Partial Credit Model pueden estar parcialmente ordenadas. En cuestionarios de CVRS los ítems suelen ser de respuestas tipo Likert y por lo tanto, son de respuesta ordenada, con lo que los modelos Graded Response Model resultan de gran utilidad, ya que estos además requieren de algo menos muestra que los Generalized Partial Credit Models (Hambleton y cols., 1991).

Entre los modelos dicotómicos los más completos son los modelos de 3–parámetros, pero los modelos de 2–parámetros también son muy útiles y sirven para ítems de actitud (attitudinal items). Aunque algunas veces también se utilizan modelos de 1–parámetro, ya que no siempre se dispone de muestra suficiente para estimar modelos más complejos. Lo mismo ocurre entre los modelos para variables politómicas, donde algunas veces no podemos considerar modelos que varíe la discriminación de los ítems (Graded Response Model o Generalized Partial Credit Model) y debemos utilizar aquellos en los que no varíe (Partial Credit Model o Rating Scale Model). Utilizar un modelo que se estima con mayor precisión algunas veces puede producir mejores resultados que utilizar modelos más complejos estimados de manera pobre, incluso aunque los modelos complejos ajusten mejor los datos (DeMars, 2010).

Aunque el uso de los modelos de la TRI tiene muchas ventajas en el desarrollo y validación de instrumentos de medida de la CVRS, no está libre de limitaciones que influyen en el lento desarrollo y adopción de esta metodología. Una de las primeras limitaciones es la muestra que se requiere para su aplicación, ya que para poder aplicar los modelos más sencillos de 1–parámetro se requieren muestras de al menos 200 respondedores, y para modelos más complejos (de 2 o más parámetros) se requieren de 2 a 5 veces ese número de respondedores (Huang y Speight, 2013; Linacre, 1994). Otra limitación es las condiciones adicionales de los modelos de la TRI, que los modelos basados en la TCT no lo requieren. La confianza de los resultados de los modelos de la TRI depende enormemente del grado en que las condiciones de los modelos se cumplen y por lo tanto es necesario realizar los test explícitos para la comprobación de dichas hipótesis. Otra limitación radica en que los modelos de la TRI asumen que las variables observables reflejan el valor de la variable latente, y que las correlaciones entre ítems surgen únicamente en virtud de la relación con la variable latente. Por tanto, queda implícito que todos los ítems son indicadores. Así, estos modelos son inapropiados para escalas de síntomas u otras variables causales. Y la última limitación es la complejidad de los modelos de la TRI, que generalmente no se pueden realizar con los paquetes estadísticos estándar y requieren de programas específicos (Huang y Speight, 2013).

Diferencias principales entre la teoría clásica del test y la teoría de la respuesta al ítem

A diferencia de la TCT, los modelos de la TRI describen matemáticamente la relación entre los individuos que están siendo medidos y los ítems que utilizamos para medir a esos individuos en la misma escala continua, lo cual nos permite establecer una dimensión subyacente en un continuo, de forma que los individuos y los ítems pueden reflejarse simultáneamente en dicho continuo, de manera ordenada, y así se pueden comparar. Esto nos permite examinar el contenido y el grado de magnitud (severidad o dificultad) de las preguntas o ítems a lo largo de dicho continuo, y así entender mejor y definir lo que estamos midiendo (Cella y cols., 2002).

Otra gran diferencia entre las dos teorías radica en que las puntuaciones de CVRS derivadas a partir de la TRI son independientes de la muestra y la escala ya que la mayor propiedad de los modelos de la TRI es la propiedad de la invarianza. Esto implica que los parámetros que caracterizan al ítem no dependen de la distribución de la habilidad de los individuos considerados, y de la misma forma, los parámetros que caracterizan a los individuos no depende del conjunto de ítems que hayamos considerado (Hambleton y cols., 1991). Teóricamente, los parámetros que estimamos sobre los ítems son invariantes; es decir, las propiedades psicométricas de los ítems no difieren según las características de los sujetos. Por el contrario, la característica de la TCT es que las propiedades del instrumento dependen de las características de los individuos. Por tanto, las puntuaciones son dependientes tanto de la muestra como de la escala (Huang y Speight, 2013).

En comparación con los modelos de la TRI, el uso de la TCT para el desarrollo y validación de instrumentos de CVRS es relativamente sencillo de entender para los clínicos, en comparación con la TRI, ya que estos últimos son metodológicamente mucho más complejos y requieren de programas específicos para la realización de los análisis (Huang y Speight, 2013). Sin embargo, una cuestión que ya entró en

debate a finales de la década de los 90 ha sido la fiabilidad de técnicas estadísticas como el análisis factorial utilizadas para la construcción de escalas de CVRS y reducción de ítems (Fayers y cols., 1997; Fayers y Hand, 1997; Juniper y cols., 1997), así como las limitaciones de los instrumentos desarrollados en base a la TCT (Embretson y Hershberger, 1999; Huang y Speight, 2013; Prieto y Delgado, 2003). Aunque la validez de la mayoría de los cuestionarios de CVRS se ha realizado mediante métodos englobados dentro de la psicometría clásica (Fayers y Machin, 2007; Gao y cols., 2004; Keller y cols., 1998; Rodgers y cols., 2005), la literatura científica actual sugiere métodos de validación más modernos, basados principalmente en la TRI (Fayers y Machin, 2007; Tennant y cols., 2004; Wright y Stone, 1979). Para mejorar las cuestiones psicométricas de la TCT, varios grupos han dedicado sus esfuerzos a la metodología de la TRI. Por ejemplo, el "National Institute of Health" lanzó en el año 2004 un proyecto de sistemas de información de medición de resultados percibidos por los pacientes (Patient-Reported Outcomes Measurement Information System) que se dedica a mejorar la fiabilidad, validez y precisión de herramientas de medida de resultados percibidos por los pacientes utilizando la TRI (Huang y Speight, 2013).

Sin embargo, el hecho de aplicar modelos de la TRI a cuestionarios que han sido diseñados y elaborados mediante la TCT también puede resultar complicado. El diseño de las escalas usando la TRI es notablemente diferente de las escalas que han sido basadas en métodos tradicionales. Las escalas generadas a través de la suma de los ítems basadas en la teoría clásica, se suelen basar en ítems con niveles de dificultad similares, y con diferentes opciones de respuesta para reflejar diferente nivel de severidad. Por el contrario, las escalas basadas en la teoría moderna o TRI se basan en ítems que representan diferentes niveles de dificultad. Así, si aplicamos un modelo de la TRI a una escala desarrollada desde la perspectiva de la TCT, nos podemos encontrar con que la escala cumple los criterios de los modelos de la TRI en lo que se refiere a unidimensionalidad, independencia local, o la no presencia de funcionamiento diferencial del ítem, pero puede resultar muy difícil encontrarnos con ítems de muy diferente nivel de dificultad. Lo más probable es que nos encontremos con ítems de similar nivel de

dificultad por el propio diseño de la escala, tal y como hemos observado en algunos ítems de la escala de función de la versión reducida del cuestionario WOMAC (Table 5.3 del Capítulo 5), así como en algunos ítems de las dos escalas resumen del cuestionario reducido HeRQoLED-S (Table 6.1, Table 6.2 y Figure 6.2 del Capítulo 6). Sin embargo no por ello vamos a desechar dichas escalas. Esos ítems serían los que consideraríamos como ítems redundantes si estaríamos diseñando el cuestionario desde la perspectiva de la TRI.

Otro problema que nos podemos encontrar en las escalas generadas a partir de la TCT es que estas escalas normalmente están generadas a partir de la suma de los ítems que la componen. Es por eso, que como los modelos de la TRI son escalas que pueden tener en cuenta la discriminación del ítem, los ítems que tienen mayor capacidad de discriminación tienen mayor peso en la escala, y los que tienen menor nivel de discriminación tienen menos peso. Esto hace que las escalas basadas en modelos de la TRI sean más fiables (DeMars, 2010). Sin embargo este problema o al menos parte de este problema podría ser solventado si las escalas de la TCT las creáramos ponderando los ítems en función de los pesos factoriales estimados. En este caso, la diferencia entre las escalas creadas mediante la TCT y las creadas mediante los modelos de la TRI quizás no sería tan notable.

8.1.2. Complementariedad de la teoría clásica del test y la teoría de la respuesta al ítem

Como hemos visto, existen grandes diferencias entre los modelos de la TCT y los modelos de la TRI. Y a su vez, entre cada una de las diferentes teorías, hemos visto como unos métodos estadísticos son más o menos adecuados dependiendo de las características de los instrumentos que estemos manejando. Sin embargo, cada una de estas perspectivas no han de ser excluyentes, es decir, no es necesario usar una o la otra, ya que el tipo de soporte que proporcionan en el proceso de evaluación de estructura es diferente con lo que el uso combinado de ambas perspectivas puede resultar de gran utilidad.

Aunque la literatura científica sugiere métodos de validación más modernos, basados principalmente en la TRI (Fayers y Machin, 2007; Tennant y cols., 2004; Wright y Stone, 1979), estos métodos no siempre son suficientes si los usamos de manera aislada. Por ejemplo, una asunción de los modelos de la TRI es la unidimensionalidad, asumiendo que los ítems de la escala son de una única dimensión o constructo. Esto implica que cuando utilizamos modelos de la TRI, estos modelos los debemos aplicar a cada una de las escalas del cuestionario de manera separada. Así, estos métodos no nos proporcionan ningún tipo de soporte acerca del modelo de medición del cuestionario al considerarlo en completo. De esta forma, es difícil confirmar una estructura multidimensional de un cuestionario de CVRS. Sin embargo, la inmensa mayoría de cuestionarios de CVRS son multidimensionales, y la necesidad de confirmar dicha estructura multidimensional es algo muy habitual y necesario en este campo. Por el contrario, mientras que los modelos de la TRI se centran en la unidimensionalidad, los modelos de la TCT se centran en la multidimensionalidad. Esta limitación de los modelos de la TRI no afecta a cuestionarios unidimensionales, como por ejemplo el cuestionario OP, pero si a cuestionarios multidimensionales, como por ejemplo el cuestionario reducido del WOMAC o el HeRQoLEDv2 y HeRQoLED-S.

Por ejemplo, en el caso de la versión reducida del cuestionario WOMAC, donde en primer lugar hemos planteado una versión reducida del cuestionario, con 11 ítems que se conforman en dos factores, dolor y función, hemos aplicado tanto métodos de la TCT como de la TRI, ya que la información que nos proporciona cada una de las técnicas es complementaria. El AFC de la TCT nos proporciona información sobre si esos 11 ítems se conforman realmente en dos factores, comprobando así la multidimensionalidad de dicho cuestionario. Mientras que el método Rating Scale Model de la TRI nos proporciona información sobre si cada una de las escalas, dolor y función, es unidimensional, y si los ítems presentan diferentes niveles de dificultad o si por el contrario hay ítems con mismo nivel de dificultad. Estas dos técnicas proporcionan información complementaria.

Otro claro ejemplo de la complementariedad de estas dos perspectivas se observa mediante los resultados obtenidos para el cuestionario HeRQoLEDv2. En este caso, en primer lugar se hipotetiza que 40 ítems se conforman en siete factores de primer orden, y que a su vez, estos se conforman en dos medidas resumen o factores de segundo orden. Para comprobar esta hipótesis es necesario utilizar modelos de la TCT, más concretamente MEE, ya que los modelos de la TRI no dan respuesta a este tipo de planteamientos. Una vez comprobado que realmente existen esas dos medidas resumen en el cuestionario HeRQoLEDv2, una de ellas formada por 19 ítems y la otra por 21 ítems, se decide reducir dicho cuestionario. Como ya hemos comentado, los instrumentos de CVRS más reducidos tienen un mejor grado de cumplimiento y una mayor tasa y calidad de respuesta, por lo que disponer de una versión reducida de un cuestionario de CVRS que mantenga las mismas buenas propiedades psicométricas que su versión larga, es de gran utilidad tanto en la práctica clínica como en investigación (Coste y cols., 1997; Moran y cols., 2001; Prieto y cols., 2003). Para el proceso de reducción de cuestionarios, si utilizamos los modelos factoriales de la TCT, el criterio que tenemos para excluir ítems ha de ser el peso factorial, de forma que aquellos ítems con menor peso factorial serán los que excluyamos de la escala. Sin embargo, en los modelos de la TRI, en primer lugar podemos excluir los ítems que no cumplen los criterios de unidimensionalidad, independencia local o la no presencia de funcionamiento diferencial del ítem. Y una vez que todos los ítems cumplen con los criterios establecidos en la TRI, estos métodos nos permiten eliminar ítems de la escala por tener mismo nivel de dificultad. De esta forma, conseguimos mantener ítems que cubran diferente nivel de dificultad para poder abarcar a todas los niveles de habilidad de los individuos. Por ejemplo, si tenemos en la escala dos ítems con el mismo nivel de discriminación (es decir, similar peso factorial) y el mismo nivel de dificultad, a la hora de reducir un cuestionario, lo más sensato sería eliminar uno de ellos, ya que con el otro mantendríamos el mismo nivel de dificultad con el mismo nivel de discriminación. Esto es imposible detectar con la TCT. Así, podemos decir que para el proceso de reducción de cuestionarios de CVRS, los métodos de la TRI son mucho más útiles que los métodos de la TCT. Nosotros hemos aplicado la TRI para la reducción de las dos medidas resumen del

cuestionario HeRQoLEDv2, eliminando los ítems redundantes con mismo nivel de dificultad. Una vez obtenida la versión reducida, el HeRQoLED-S, hemos vuelto a aplicar métodos tanto de la TCT como de la TRI para revalidar el cuestionario reducido en una nueva muestra de pacientes, de forma que la TCT nos ha proporcionado información sobre la multidimensionalidad de la versión reducida, algo imposible de comprobar mediante la TRI, mientras que la TRI nos ha proporcionado información sobre la unidimensionalidad de cada una de las dos medidas resumen, aportando información de la capacidad de discriminación de cada ítem dentro de cada escala.

En definitiva, el soporte que nos proporcionan los métodos de la TCT y los de la TRI a la hora de evaluar la validez de estructura de cuestionarios de CVRS es diferente aunque complementaria. Los métodos de la TCT nos proporcionan información sobre la dimensionalidad, la homogeneidad y el solapamiento entre variables latentes de un cuestionario. Es decir, nos proporcionan información sobre los siguientes aspectos: a) si todos los ítems de una escala se relacionan con una variable latente o constructo, o si existe evidencia de que se necesitan más variables latentes para explicar la variabilidad observada en los ítems; b) si todos los ítems de una subescala parecen tener el mismo peso sobre la variable latente; o c) sobre si algunos ítems de una subescala correlacionan con otras variables latentes. Así, los métodos de la TCT analizan el cuestionario como un total, es decir, con todos sus ítems y todos sus constructos o factores, proporcionando información sobre si el modelo de medición considerado es adecuado o no. Por otro lado, los métodos de la TRI se centran principalmente en la unidimensionalidad, y cada escala o constructo del cuestionario se analiza de manera separada. Estos métodos nos informan sobre si los ítems que componen una escala son unidimensionales, nos informan del grado de dificultad que presenta cada ítem, nos permite ordenar a su vez estos ítems en función del nivel de dificultad, y nos permite obtener el nivel de discriminación de cada uno de los ítems. Por tanto, mientras una de las perspectivas nos proporciona soporte sobre la dimensionalidad, la otra nos proporciona soporte sobre la unidimensionalidad.

Así, el uso combinado de ambas perspectivas resulta de gran utilidad, proporcionando una mayor evidencia de la validez de un instrumento de CVRS.

Otro aspecto en el que las dos perspectivas combinadas resultan de gran utilidad es en la elección del modelo de la TRI a utilizar. Como hemos dicho, en los modelos de 1–parámetro se asume que el parámetro de discriminación es constante para todos los ítems. Es decir, se asume que todos los ítems tienen la misma capacidad de discriminación. El parámetro de discriminación se puede asemejar al peso factorial que obtenemos de los modelos de la TCT. A través del peso factorial sabemos que ítems aportan más a ese factor y que ítems aportan menos. Así, los pesos factoriales obtenidos mediante la TCT pueden ser de gran ayuda a la hora de decidir si aplicamos un método de la TRI de 1–parámetro o de 2–parámetros. Esto resulta de especial utilidad cuando la muestra disponible no es muy grande. En muchas ocasiones nos podemos encontrar con la situación de querer aplicar modelos de 2–parámetros por el simple hecho de ser más completos, pero sin embargo, la muestra disponible suele resultar una limitación. En estos casos, si por ejemplo los pesos factoriales obtenidos mediante los métodos de la TCT son similares, ello nos proporciona de alguna manera evidencia de la no necesidad de considerar modelos de 2–parámetros. En definitiva, el uso combinado de ambas perspectivas resulta de gran utilidad en el proceso de evaluación de la validez de estructura de instrumentos de CVRS.

8.1.3. Recomendaciones

Teniendo en cuenta las ventajas y limitaciones de cada una de las dos perspectivas de enfoque, la TCT y la TRI, así como las ventajas y limitaciones de cada uno de los métodos dentro de cada teoría, proporcionaremos algunas recomendaciones a la hora de validar y/o reducir cuestionarios de CVRS.

- Se recomienda combinar el uso de métodos de la TCT con métodos de la TRI.

- Entre los métodos de la TCT, no se recomienda el uso del AFE de manera aislada. Siempre debiera ir acompañado del AFC o MEE, según lo requiera la estructura del cuestionario.
- Solo se recomienda el uso del AFE en caso de desconocimiento de la estructura subyacente del cuestionario.
- En caso de estar validando la adaptación y traducción de un cuestionario a otra cultura y lengua, o la reducción de un cuestionario, en cuyo caso el modelo de medición del instrumento es bien sabido, entre los métodos de la TCT siempre debemos utilizar el AFC o MEE, según lo requiera la estructura del cuestionario.
- Entre los métodos de la TCT, en caso de que el cuestionario tenga preguntas con respuestas dicotómicas, o politómicas pero de menos de 5 categorías de respuesta, se recomienda la utilización del AFC o los MEE específicos para datos categóricos.
- Entre los métodos de la TCT, en caso de que el cuestionario tenga preguntas con respuestas de 5 categorías o más, solo se recomienda la utilización del AFC y MEE clásicos siempre y cuando la muestra sea suficientemente grande y los datos no sean extremadamente sesgados. En caso contrario, se recomienda utilizar el AFC o MEE específico para datos categóricos.
- Entre los métodos de la TRI, se recomienda utilizar los métodos de 2–parámetros, ante los de 1–parámetro, siempre y cuando la muestra sea suficientemente grande, y los pesos factoriales obtenidos mediante métodos de la TCT sean muy diferentes entre los ítems de una escala.
- Entre los métodos de la TRI, en caso de no disponer de suficiente muestra y siempre que los pesos factoriales obtenidos mediante métodos de la TCT sean similares entre los ítems de una escala, se recomienda utilizar los métodos de 1–parámetro.
- Entre los métodos de 1–parámetro para datos politómicos dentro de la TRI, se recomienda utilizar el Rating Scale Model siempre que los ítems de la escala tengan el mismo número de opciones de respuesta.

- Entre los métodos de 2–parámetros para datos politómicos dentro de la TRI, se recomienda utilizar el Graded Response Model siempre que los ítems de la escala sean de respuesta ordinal.
- En caso de que en la estructura del cuestionario tengamos variables causales, solo debemos utilizar los MEE.
- Para la reducción de cuestionarios se recomienda utilizar modelos de la TRI.

8.2. Propiedades psicométricas de los tres cuestionarios de calidad de vida relacionada con la salud evaluados

Como ya hemos comentado, es fundamental disponer de instrumentos de medición de la salud y de la CVRS evaluados debidamente y con buenas propiedades psicométricas. Así, en esta sección discutiremos brevemente los resultados psicométricos de fiabilidad, validez y sensibilidad al cambio obtenidos para cada uno de los cuestionarios evaluados en la presente tesis, los cuestionarios OP, WOMAC, y HeRQoLEDv2, así como su versión reducida. Para ello tendremos en cuenta los criterios que se muestran resumidos en la Figura 1.1 del Capítulo 1, establecidos por la SAC (Aaronson y cols., 2002) y apoyados por otros muchos grupos (Huang y Speight, 2013), centrándonos en los aspectos que requieren del análisis numérico y estadístico, y más concretamente en la evaluación de la validez de estructura.

Propiedades psicométricas del cuestionario OP

El cuestionario específico OP (Karlsson y cols., 2003; Sullivan y cols., 1993), diseñado para medir el impacto de la obesidad mórbida en el funcionamiento psicosocial, consta de ocho ítems y se conforman en un único factor o constructo, presentando así una estructura sencilla (Figura 1.3). Tal y como se muestra en el Capítulo 4, la adaptación cultural y traducción del cuestionario OP al español se ha realizado siguiendo una rigurosa metodología. En cuanto a sus propiedades psicométricas, queda demostrada su fiabilidad a través del estudio de la

consistencia interna. A pesar de no disponer de un estudio de reproducibilidad, se ha estudiado la homogeneidad de los ítems mediante el estudio de validez convergente y discriminante de los ítems. Por otro lado, queda demostrada la validez convergente y discriminante de las escalas, así como la validez de grupos conocidos. La validez de estructura de este cuestionario se ha estudiado mediante la aplicación únicamente de métodos de la TCT, el AFE y el AFC. Tal y como hemos explicado en la sección anterior, no se recomienda utilizar el AFE de manera aislada, y menos aun cuando estamos estudiando la validez de estructura de un cuestionario adaptado y traducido a otra lengua, como es este caso. Para solventar esta deficiencia, hemos acompañado el AFE con un AFC, proporcionando así mayor evidencia de la validez de estructura del instrumento. En cuanto al AFC, se ha utilizado el AFC clásico y no específico para datos categóricos a pesar de que los ítems solo tenían 4 opciones de respuesta. Una de las recomendaciones es la utilización de AFC para datos categóricos siempre y cuando los ítems no tengan 5 o más opciones de respuesta, ya que se ha demostrado que estas técnicas son robustas en estos casos para muestras suficientemente grandes (Cohen y cols., 2003). Sin embargo, en nuestro caso, la muestra no era tan grande, 123 pacientes. Este tamaño muestral es una limitación a la hora de poder aplicar técnicas como el AFC para datos categóricos ya que requieren de muestras mayores. La muestra reducida también supone una limitación para la utilización de modelos de la TRI. Así, la utilización de modelos de la TRI en una muestra mayor de pacientes con obesidad mórbida evidenciaría aún más la validez de estructura de este cuestionario. Por otro lado, falta estudiar la sensibilidad al cambio del instrumento, para garantizar su utilidad en estudios longitudinales.

En conclusión, podemos decir que este cuestionario es fiable y valido. Aunque más análisis evidenciarían mejor su validez, nuestros resultados confirman que este cuestionario ha sido adaptado y traducido al español de manera correcta y siguiendo una metodología rigurosa. Además, al menos parcialmente, cumple los requisitos psicométricos que se requieren para este tipo de cuestionarios, proporcionando así un instrumento para medir el impacto de la obesidad mórbida

en el funcionamiento psicosocial, que puede ser utilizado entre investigadores, clínicos y decisores sanitarios.

Propiedades psicométricas de la versión reducida propuesta del cuestionario WOMAC

La versión reducida propuesta del cuestionario WOMAC (Davis y cols., 2003; Tubach y cols., 2005), es un cuestionario específico empleado para la medición de sintomatología y función en pacientes con artrosis de cadera. Este consta de 11 ítems que se conforman en dos factores o constructos, dolor y función, presentando así una estructura multidimensional pero sencilla (Figura 1.6). Tal y como se muestra en el Capítulo 5 de la presente tesis, esta versión reducida ha demostrado ser fiable, valida y sensible al cambio. En cuanto a la fiabilidad, esta queda demostrada a través del estudio de la consistencia interna, a falta de un estudio de reproducibilidad. En cuanto a la validez de constructo, queda demostrada la validez convergente y discriminante, así como la validez de grupos conocidos. La validez de estructura de este cuestionario se ha estudiado mediante la combinación de métodos de la TCT y métodos de la TRI, siendo este hecho enriquecedor y proporcionando mayor evidencia sobre la validez de estructura. Finalmente, también queda demostrada la sensibilidad al cambio del instrumento a los seis meses de la artroplastia total de cadera.

Entre los métodos de la TCT, hemos empleado el AFC ya que dada la estructura multidimensional y de primer orden del cuestionario, este método era el adecuado. Además, en este caso, al estar ante la reducción de un cuestionario con una estructura subyacente bien conocida como es el WOMAC (Bellamy y cols., 1988a y 1988b), la utilización de métodos confirmatorios es fundamental. Se ha utilizado el AFC clásico en vez del específico para datos categóricos. Este cuestionario consta de preguntas politómicas con cinco opciones de respuesta, con lo que es sabido que para este tipo de ítems y teniendo en cuenta que disponíamos de una muestra muy grande, estas técnicas son robustas (Cohen y cols., 2003). Con el fin de proporcionar aun mayor evidencia de la validez de estructura del instrumento, se

ha utilizado el AFC en dos muestras de pacientes de forma separada, y ha quedado demostrado que prácticamente los resultados son los mismos. Además, esta evidencia quedó complementada con el análisis de la TRI, también aplicada a ambas muestras de pacientes. La muestra que disponíamos en este caso para la aplicación de los modelos de la TRI no suponía ninguna limitación, más aun cuando una de las escalas constaba únicamente de tres ítems y la otra de ocho ítems. En cuanto a los métodos de la TRI se aplicó el Rating Scale Model a cada una de las escalas de la versión reducida del cuestionario WOMAC. Todas las preguntas del cuestionario disponen del mismo número de opciones de respuesta (Anexo IV), con lo que este método es adecuado dadas las características del cuestionario. El modelo utilizado es un modelo de la TRI de 1–parámetro asumiendo que el parámetro de discriminación es constante para todos los ítems. Los pesos factoriales que se obtuvieron del AFC no eran tan similares entre sí, lo que la utilización de modelos de 2–parámetros podría ser una asignatura pendiente por realizar. A pesar de ello, los resultados de los modelos de la TRI fueron muy similares en ambas muestras de pacientes proporcionando mayor evidencia de su validez de estructura y estabilidad.

En conclusión, podemos decir que este cuestionario es fiable, valido y sensible al cambio. Nuestros resultados confirman que este cuestionario cumple con los requisitos psicométricos que se requieren para los cuestionarios de CVRS, combinando métodos clásicos de la TCT, así como métodos modernos de la TRI, proporcionando mayor evidencia de la validez. Por tanto, esta versión reducida de 11 ítems del cuestionario WOMAC, es un instrumento simple, corto y de fácil aplicación, que sirve para medir dolor y función en pacientes con artrosis de cadera, y que puede ser utilizado con total garantía entre investigadores, clínicos y decisores sanitarios.

Confirmación de la estructura de segundo orden del cuestionario HeRQoLEDv2, reducción del cuestionario, y propiedades psicométricas del cuestionario reducido HeRQoLED-S

El cuestionario HeRQoLEDv2 (Las Hayas y cols., 2006 y 2007), específico para medir la CVRS en pacientes con trastorno de la conducta de la alimentación, consta de 55 ítems que se conforman en nueve factores, denominados factores de primer orden (Figura 1.5). A su vez, se ha demostrado que existen dos medidas resumen (denominadas factores de segundo orden) englobando siete de los nueve factores de primer orden (Figura 1.7), tal y como se muestra en el Capítulo 6 de la presente tesis. Para comprobar esta estructura interna de segundo orden en el cuestionario solo hemos empleado métodos de la TCT, pero en este caso el hecho de no haber combinado métodos de ambas teorías o perspectivas, no es una limitación, ya que los métodos de la TRI no abordan este tipo de estructura de segundo orden donde lo que precisamente queremos comprobar es la dimensionalidad del cuestionario. Concretamente, utilizamos los MEE clásicos, y no los específicos para datos categóricos. La estructura a confirmar era una estructura de segundo orden, y por lo tanto no era una estructura sencilla. El número de ítems del que partíamos era bastante elevado (40 ítems), y la muestra de la que disponíamos aunque no era pequeña dada la patología, sí que era escasa para la utilización de métodos específicos para datos categóricos, ya que estos requieren de mayor muestra (Fayers y Machin, 2007; Muthén y Muthén, 2002). Además, los ítems son de seis opciones de respuesta tipo Likert, con lo que teniendo en cuenta las puntualizaciones anteriormente citadas y que estos métodos son robustos para este número de opciones de respuesta (Cohen y cols., 2003), la utilización de MEE clásicos no creemos que suponga ninguna limitación.

El Capítulo 6 de la presente tesis doctoral muestra como la reducción del cuestionario HeRQoLEDv2 se ha realizado siguiendo métodos de la TRI, tal y como se ha recomendado, y siguiendo una rigurosa metodología. Así, de los 40 ítems que se conformaban en siete factores de primer orden, para así agruparse en dos medidas resumen, se ha pasado a un cuestionario reducido de 20 ítems, el

HeRQoLED-S (Figura 1.8). El modelo de la TRI empleado para la reducción del cuestionario fue el Rating Scale Model. Todos los ítems de cada una de las dos medidas resumen tienen el mismo número de opciones de respuesta (Anexo III), con lo que este método se adecuaba bien a la estructura. A pesar de que los pesos factoriales obtenidos mediante los MEE no eran similares en todos los casos, se utilizó este método de 1–parámetro, asumiendo que todos los ítems tenían la misma capacidad de discriminación. Sin embargo, dada la muestra de la que disponíamos y teniendo en cuenta que una de las escalas tenía 19 ítems, y la otra 21, se consideró más adecuado utilizar modelos de 1–parámetro.

Los análisis preliminares acerca de las propiedades psicométricas de esta versión reducida, mostradas en el Capítulo 6, demostraron que este cuestionario era fiable y valido. La fiabilidad quedó demostrada mediante los buenos resultados del estudio de la consistencia interna de las dos medidas resumen, a falta de un estudio de reproducibilidad. En cuanto a la validez de constructo, queda demostrada la validez convergente y discriminante, así como la validez de estructura. Para la validez de estructura, además del Rating Scale Model anteriormente mencionado, se utilizó los MEE de la TCT. Al igual que antes, en este caso también se aplicó el método clásico en vez del específico para datos categóricos, ya que a pesar de partir de la mitad de numero de ítems, la estructura seguía siendo compleja y la muestra disponible era aún más pequeña, con lo que dado que los ítems tenían seis opciones de respuesta se consideró adecuado utilizar estos métodos clásicos. Por último, en este estudio preliminar quedó sin demostrar la sensibilidad al cambio del instrumento.

Aunque el estudio preliminar sobre las propiedades psicométricas de este cuestionario mostró buenos resultados acerca de su fiabilidad y validez, este presentaba algunas limitaciones: a) dentro de la TCT se utilizó el MEE clásico en vez del específico para datos categóricos; b) dentro de la TRI se utilizaron modelos de 1–parámetro en vez de 2–parámetros; y c) con el fin de evitar estudiar la validez y fiabilidad del cuestionario reducido con la misma muestra que la propia derivación del instrumento, se utilizaron los datos de seguimiento a un año de

estos pacientes, siendo esta muestra algo más reducida. El hecho de disponer de una muestra tan reducida dificultaba la utilización de métodos más adecuados como son los modelos factoriales específicos para datos categóricos de la TCT, o los modelos de 2–parámetros de la TRI, ya que estos métodos requieren de muestras mayores (Fayers y Machin, 2007; Huang y Speight, 2013; Linacre, 1994; Muthén y Muthén, 2002).

Sin embargo, todas estas limitaciones quedan solventadas tras la revalidación del instrumento reducido HeRQoLED-S en una nueva muestra de pacientes, tal y como se muestra en el Capítulo 7 de la presente tesis. Así, queda demostrado que este cuestionario es fiable, valido y sensible al cambio. En cuanto a la fiabilidad, esta queda demostrada a través del estudio de la consistencia interna, a falta de un estudio de reproducibilidad. En cuanto a la validez de constructo, queda demostrada la validez convergente y discriminante, así como la validez de grupos conocidos. La validez de estructura se ha estudiado mediante la combinación de métodos de la TCT y métodos de la TRI, siendo este hecho enriquecedor y proporcionando mayor evidencia sobre la validez de estructura. A este hecho hay que añadir que se han utilizado métodos más adecuados dadas las características del instrumento que los utilizados en el análisis preliminar presentado en el Capítulo 6. Finalmente, también queda demostrada una moderada capacidad de sensibilidad al cambio al año de seguimiento.

En lo que se refiere a la validez de estructura, entre los métodos de la TCT, hemos empleado los MEE específicos para datos categóricos, ya que en este caso a pesar de que los ítems tienen seis opciones de respuesta, dado que la muestra era algo mayor que la anterior y que los ítems de los que partíamos ya no eran 40 sino la mitad, se aplicó este método específico con el fin de mejorar la validación preliminar. Los buenos resultados obtenidos de este análisis demuestran la estructura subyacente de este cuestionario reducido con las dos medidas resumen. Además, esta evidencia quedo complementada con el análisis de la TRI aplicada a cada una de las medidas resumen. En este caso también, con el fin de mejorar los métodos utilizados en el análisis preliminar, y dado que se disponía de una

muestra mayor, se utilizó el Graded Response Model, de 2–parámetros. Teniendo en cuenta que los pesos factoriales que se obtienen de los MEE no son similares entre sí, este método de 2–parámetros es más adecuado que el de 1–parámetro. Además, como todos los ítems del cuestionario son ordinales, este es un método adecuado de entre los de 2–parámetros para datos politómicos. Los buenos resultados obtenidos de este análisis proporcionan aún mayor evidencia de la validez de estructura del instrumento, aumentando además la precisión en la estimación de los parámetros que lo caracterizan.

En conclusión, podemos decir que este cuestionario es fiable, valido y sensible al cambio. Nuestros resultados confirman que este cuestionario cumple con los requisitos psicométricos que se requieren para los cuestionarios de CVRS, además de que la combinación de métodos clásicos de la TCT, así como métodos modernos de la TRI, proporcionan mayor evidencia de la validez. Por tanto, esta versión reducida de 20 ítems, es un instrumento simple, corto y de fácil aplicación, que sirve para medir aspectos como la adaptabilidad social y la salud mental y el rendimiento en pacientes con un trastorno de la conducta de la alimentación. El completo estudio de validación realizado y el uso de metodología variada y adecuada hacen que este se pueda utilizar con total garantía. Además, su simplicidad y aplicabilidad, hacen que aumente su aceptabilidad tanto entre investigadores, clínicos o decisores sanitarios, siendo así una buena alternativa a la versión larga del cuestionario.

8.3. Líneas de investigación futura

En este trabajo hemos presentado las ventajas y limitaciones de algunos métodos de la TCT, así como algunos métodos de la TRI, y la complementariedad entre ellos, a la hora de validar cuestionarios de CVRS, planteando así algunas recomendaciones de actuación. Por otro lado, hemos aplicado estos métodos de validación y/o reducción a los tres cuestionarios específicos de CVRS de tres ámbitos distintos de la salud, patología endocrinológica, osteomuscular, y psiquiátrica, demostrando sus buenos resultados psicométricos que garantizan su

uso entre investigadores, clínicos y decisores sanitarios. Sin embargo, estas conclusiones constituyen relevantes puntos de partida y no de llegada.

Por ello, en lo que se refiere a los cuestionarios utilizados, quedaría por estudiar aquellos aspectos psicométricos de los diferentes cuestionarios utilizados que han quedado sin analizar en este trabajo. Así, planteamos las siguientes líneas de investigación a futuro:

- En el caso del cuestionario específico OP diseñado para medir el impacto de la obesidad mórbida en el funcionamiento psicosocial, cabría la posibilidad de utilizar el AFC específico para datos categóricos así como modelos de la TRI en una muestra mayor de pacientes, para complementar los resultados de validez de este instrumento y poder compararlos con los obtenidos en este trabajo. Además, sería necesario realizar un estudio de la sensibilidad al cambio, ya que esta es una de las propiedades psicométricas que ha quedado sin estudiar.
- En el caso de la versión reducida propuesta del cuestionario WOMAC, empleado para la medición de dolor y función en pacientes con artrosis de cadera, sería interesante utilizar el AFC específico para datos categóricos así como modelos de la TRI de 2–parámetros, para complementar los resultados de validez de este instrumento y compararlos con los ya obtenidos en este trabajo. Por otro lado, en lo que se refiere a este instrumento, también sería interesante validarlo en pacientes con artrosis de rodilla.
- En el caso del cuestionario específico HeRQoLED-S, diseñado para medir la CVRS en pacientes con un trastorno de la conducta de la alimentación, sería necesario realizar un estudio de la sensibilidad al cambio en una muestra de pacientes mayor.

Por otro lado, en cuanto a los métodos estadísticos para la evaluación de la validez de estructura, aunque en este trabajo hemos descrito, aplicado y discutido un amplio abanico de métodos, aún existen más métodos, quizás algo menos desarrollados o utilizados en este ámbito. Por ello, planteamos las siguientes líneas de investigación a futuro:

- Para futuros estudios, sería interesante utilizar el análisis Mokken en el proceso de validación de cuestionarios de CVRS. Los modelos Mokken son modelos no paramétricos de la TRI en los que no se asume ninguna distribución de probabilidad subyacente al rasgo latente (Sijtsma y Molenaar, 2002).
- Otra línea de investigación en la que trabajar sería en los modelos de la TRI multidimensionales. Estos modelos multidimensionales fueron originalmente introducidos por Lord y Novik (1968) y Samejima (1974), y posteriormente por Embretson (1984) y McDonald (1989). Los cuestionarios de CVRS son generalmente instrumentos multidimensionales con lo que estos modelos resultarían de gran utilidad en este ámbito.
- Otra línea de investigación en la que trabajar es la utilización del Análisis de Correspondencias Múltiples en el proceso de validación de instrumentos de CVRS. Se trata de un método exploratorio cuyo objetivo es similar al del análisis factorial, es decir, resumir una cantidad de datos (ítems) en un número reducido de dimensiones o constructos. Los ítems vienen representados a través de "mapas", que pueden ser de utilidad para explorar relaciones entre los ítems (Benzécri, 1973; Greenacre, 1984). Aunque este método fue desarrollado hace tiempo, no ha sido muy utilizado en el campo de los cuestionarios de CVRS. Esta técnica es específica para datos categóricos, lo que la hace interesante en este ámbito.

Capítulo 9

Conclusiones

En vista a los resultados obtenidos en este trabajo, se han obtenido las siguientes conclusiones:

1. Los métodos estadísticos basados en la teoría clásica del test (TCT) y los basados en la teoría de la respuesta al ítem (TRI) son complementarios entre sí. Mientras que los primeros nos proporcionan información sobre la dimensionalidad, la homogeneidad y el solapamiento entre variables latentes del cuestionario, los segundos se centran en la unidimensionalidad. Así, cada uno de los métodos proporciona diferente tipo de soporte en el proceso de validación y/o reducción de cuestionarios de calidad de vida relacionada con la salud (CVRS).

   Los métodos estadísticos basados en la TCT pueden ayudar en la elección del modelo de la TRI, en el proceso de la validación de estructura de un

cuestionario de CVRS. La similitud o diferencia entre los pesos factoriales obtenidos mediante los métodos de la TCT ayudan en la elección de utilizar modelos de 1 o 2–parámetros de la TRI.

2. La combinación de diferentes métodos estadísticos, tanto de la TCT como de la TRI, en el proceso de validación y/o reducción de cuestionarios de CVRS, proporciona mayor evidencia de su validez y de sus buenas propiedades psicométricas, confirmando la dimensionalidad del instrumento y la unidimensionalidad de cada escala del instrumento. Ello hace que la utilización combinada de estas técnicas estadísticas proporcione versiones científicamente validadas en castellano de cuestionarios de CVRS, para diferentes ámbitos de la salud, que pueden ser utilizadas en la práctica clínica, así como entre investigadores y decisores sanitarios con total garantía.

3. La adaptación cultural y traducción del cuestionario Obesity-related Problems scale (OP) al español ha sido realizada siguiendo una rigurosa metodología. Y los métodos estadísticos basados en la TCT han permitido comprobar su fiabilidad y validez. Así, a pesar de ser necesario estudiar la sensibilidad al cambio de este cuestionario, los resultados obtenidos demuestran que el cuestionario presenta buenas propiedades psicométricas de validez y fiabilidad, proporcionando así un instrumento para medir el impacto de la obesidad mórbida en el funcionamiento psicosocial, que puede utilizado con total garantía al menos en estudios transversales.

4. Los métodos estadísticos de la TCT así como de la TRI han demostrado que la versión reducida propuesta del cuestionario específico Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) es fiable, valido y sensible al cambio. Además, la combinación de métodos de la teoría clásica y la moderna proporcionan mayor evidencia sobre su validez y buenas propiedades psicométricas. Así, los resultados obtenidos demuestran que esta versión reducida puede ser utilizada con total garantía para la medición de dolor y función en pacientes con artrosis de cadera.

5. Los Modelos de Ecuaciones Estructurales de la TCT han permitido demostrar que el cuestionario Health Related Quality of Life for Eating Disorders version 2 (HeRQoLEDv2), diseñado para medir la CVRS en pacientes con un trastorno de la conducta de la alimentación, tiene una estructura interna subyacente de segundo orden. Así, siete de los factores de primer orden de dicho cuestionario se resumen a su vez en dos medidas resumen, denominadas "Adaptabilidad social" y "Salud mental y rendimiento".

6. La utilización de métodos de la TRI han permitido reducir el cuestionario específico HeRQoLEDv2, dando así lugar a la versión reducida Health Related Quality of Life for Eating Disorders–Short Version (HeRQoLED-S) de solo 20 ítems, proporcionando una herramienta simple, corta y de fácil aplicación.

7. Los métodos estadísticos de la TCT así como de la TRI han demostrado que la versión reducida HeRQoLED-S es fiable, valida y sensible al cambio, en una patología en la que el cambio esperado es pequeño y el tiempo estimado de recuperación es mayor de un año. La combinación de métodos de la teoría clásica y la moderna, además de la utilización de diferentes métodos dentro de cada una de las dos teorías, proporciona mayor evidencia sobre su validez de estructura y buenas propiedades psicométricas. Así, los resultados obtenidos demuestran que esta versión reducida y científicamente validada en castellano puede ser utilizada para la medición de la CVRS en pacientes con un trastorno de la conducta de la alimentación.

Bibliografía

[1] Aaronson N, Alonso J, Burnam A, Lohr KN, Patrick DL, Perrin E, et al. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res. 2002;11(3):193-205.

[2] Aaronson NK, Acquadro C, Alonso J, Apolone G, Bucquet D, Bullinger M, et al. International Quality of Life Assessment (IQOLA) Project. Qual Life Res. 1992;1(5):349-51.

[3] Abraham SF, Brown T, Boyd C, Luscombe G, Russell J. Quality of life: eating disorders. Aust N Z J Psychiatry. 2006;40(2):150-5.

[4] Adair CE, Marcoux GC, Cram BS, Ewashen CJ, Chafe J, Cassin SE, et al. Development and multi-site validation of a new condition-specific quality of life measure for eating disorders. Health Qual Life Outcomes. 2007;5:23.

[5] Akaike H. A new look at the statistical model identification. IEE Transactions on Automatic Control. 1974;19(6):716-23.

[6] Alonso J, Prieto L, Anto JM. The Spanish version of the SF-36 Health Survey (the SF-36 health questionnaire): an instrument for measuring clinical results. Med Clin (Barc). 1995;104(20):771-6.

[7] American Educational Research Assocation, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Psychological Association; 1999.

[8] American Psychiatric Association. Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: American Psychiatric Publishing, Inc.; 1994.

[9] Anderson JG, Wixson RL, Tsai D, Stulberg SD, Chang RW. Functional outcome and patient satisfaction in total knee patients over the age of 75. J Arthroplasty. 1996;11(7):831-40.

[10] Andrich D. Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement. 1978;2(4):581-94.

[11] Anitua C, Aizpuru F, Sanzo JM. Encuesta de Salud 1997. Mejorando la Salud. Vitoria: Departamento de Sanidad. Gobierno Vasco; 1997.

[12] Aranceta J, Perez RC, Serra ML, Ribas BL, Quiles IJ, Vioque J, et al. Prevalence of obesity in Spain: results of the SEEDO 2000 study. Med Clin (Barc). 2003;120(16):608-12.

[13] Aranda FF, Villar MB, Murcia SJ, Gil VT, Ruiloba JV, Vilches IG. Outpatient group psychotherapy for anorexia nervosa. Anales de Psiquiatría. 1997;13(6):236-42.

[14] Arrindell WA, Van der Ende J. An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. Applied Psychological Measurement. 1985;9(2):165-78.

[15] Asparouhov T, Muthén BO. Bayesian analysis of latent variable models using Mplus. Technical Report. Version 4; 2010.

[16] Badia X, Alonso J. La medida de la salud. Guías de escala de medición en español (4th ed.). Barcelona: Tecnología y Ediciones del Conocimiento (EDITEC); 2007.

[17] Badia X, Roset M, Montserrat S, Herdman M, Segura A. The Spanish version of EuroQol: a description and its applications. European Quality of Life scale. Med Clin (Barc). 1999;112 Suppl 1:79-85.

[18] Badia X, Schiaffino A, Alonso J, Herdman M. Using the EuroQoI 5-D in the Catalan general population: feasibility and construct validity. Qual Life Res. 1998;7(4):311-22.

[19] Bamford BH. Assessing quality of life in the eating disorders: the HeRQoLED-S. Expert Rev Pharmacoecon Outcomes Res. 2010;10(5):513-6.

[20] Bardone-Cone AM, Abramson LY, Vohs KD, Heatherton TF, Joiner TE, Jr. Predicting bulimic symptoms: an interactive model of self-efficacy, perfectionism, and perceived weight status. Behav Res Ther. 2006;44(1):27-42.

[21] Bardone-Cone AM, Joiner TE, Jr., Crosby RD, Crow SJ, Klein MH, le Grange D, et al. Examining a psychosocial interactive model of binge eating and vomiting in women with bulimia nervosa and subthreshold bulimia nervosa. Behav Res Ther. 2008;46(7):887-94.

[22] Baron G, Tubach F, Ravaud P, Logeart I, Dougados M. Validation of a short form of the Western Ontario and McMaster Universities Osteoarthritis Index function subscale in hip and knee osteoarthritis. Arthritis Rheum. 2007;57(4):633-8.

[23] Bartholomew DJ. Three faces of factor analysis. In: Cudeck R, MacCallum RC, editors. Factor analysis at 100. Historical development and future directions. Mahwah, NJ: Lawrence Erlbaum Associates Publishers; 2007. p. 9-21.

[24] Bartholomew DJ, Steele F, Moustaki I, Galbrain JI. The analysis and interpretation of multivariate data for social scientists. Boca Raton, FL: Chapman & Hall/CRC; 2002.

[25] Basu D. Quality-Of-Life Issues in Mental Health Care: Past, Present, and Future. German Journal of Psychiatry. 2004;7(3):35-43.

[26] Batista-Foguet JM, Coenders G, Alonso J. Confirmatory factor analysis. Its role on the validation of health related questionnaires. Med Clin (Barc). 2004;122 Suppl 1:21-7.

[27] Beales DL, Dolton R. Eating disordered patients: personality, alexithymia, and implications for primary care. Br J Gen Pract. 2000;50(450):21-6.

[28] Beaton DE, Wright JG, Katz JN. Development of the QuickDASH: comparison of three item-reduction approaches. J Bone Joint Surg Am. 2005;87(5):1038-46.

[29] Bellamy N. WOMAC osteoarthritis index: a user's guide, IV. London: 2000.

[30] Bellamy N, Buchanan WW, Goldsmith CH, Campbell J. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes following total hip or knee arthroplasty in osteoarthritis. J Orthop Rheumatol. 1988a;1:95-108.

[31] Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol. 1988b;15(12):1833-40.

[32] Benzécri JP. Analyse des données. Tome I: Analyse des correspondances. Tome II: La Classification. Paris: Dunod; 1973.

[33] Bilbao A, Las Hayas C, Forero CG, Padierna A, Martin J, Quintana JM. Cross-validation study using item response theory: the health-related quality of life for eating disorders questionnaire-short version. Assessment. 2014;21(4):477-93.

[34] Bilbao A, Mar J, Mar B, Arrospide A, Martinez de Aragón G, Quintana JM. Validation of the Spanish translation of the questionnaire for the obesity-related problems scale. Obes Surg. 2009;19(10):1393-400.

[35] Bilbao A, Quintana JM, Escobar A, Las Hayas C, Orive M. Validation of a proposed WOMAC short form for patients with hip osteoarthritis. Health Qual Life Outcomes. 2011;9:75.

[36] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307-10.

[37] Bock RD. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika. 1972;37:29-51.

[38] Bohn K, Doll HA, Cooper Z, O'Connor M, Palmer RL, Fairburn CG. The measurement of impairment due to eating disorder psychopathology. Behav Res Ther. 2008;46(10):1105-10.

[39] Boini S, Briancon S, Guillemin F, Galan P, Hercberg S. Impact of cancer occurrence on health-related quality of life: a longitudinal pre-post assessment. Health Qual Life Outcomes. 2004;2:4.

[40] Bollen KA. Structural equations with latent variables. New York: John Wiley & Sons; 1989.

[41] Bond TG, Fox CM. Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.). New Jersey: Lawrence Erlbaum Associates Publishers; 2007.

[42] Bowers WA, Ansher LS. The effectiveness of cognitive behavioral therapy on changing eating disorder symptoms and psychopathology of 32 anorexia nervosa patients at hospital discharge and one year follow-up. Ann Clin Psychiatry. 2008;20(2):79-86.

[43] Bowling A. Health related quality of life: conceptual meaning, use and measurement. Measuring disease (2nd ed.). Philadelphia, PA, 19106 USA: Open University Press; 2001. p. 1-19.

[44] Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002;21(2):271-92.

[45] Browne M, Cudeck R. Alternative ways of assessing model fit. SMR. 1992;21:230-58.

[46] Butler GS, Vallis TM, Perey B, Veldhuyzen van Zanten SJ, MacDonald AS, Konok G. The Obesity Adjustment Survey: development of a scale to assess psychological adjustment to morbid obesity. Int J Obes Relat Metab Disord. 1999;23(5):505-11.

[47] Calvete E, Estevez A, Lopez de Arroyabe E, Ruiz P. The Schema Questionnaire--Short Form: Structure and Relationship with Automatic Thoughts and Symptoms of Affective Disorders. European Journal of Psychological Assessment. 2005;21(2):90-9.

[48] Castro J, Toro J, Salamero M, Guimera E. The Eating Attitudes Test: Validation of the Spanish version. Evaluación Psicológica/Psychological Assessment. 1991;7(2):175-89.

[49] Cella D, Chang C, Heinemann AW. Item Response Theory (IRT): Applications in Quality of Life Measurement, Analysis and Interpretation. Statistical Methods for Quality of Life Studies: Design, Measurements and Analysis. Netherland: Kluwer Academic Publishers; 2002.

[50] Clancy CM, Eisenberg JM. Outcomes research: measuring the end results of health care. Science. 1998;282:245-6.

[51] Cohen J. A power primer. Psychol Bull. 1992;112(1):155-9.

[52] Cohen J, Cohen P, West SG, Aiken LS. Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates Publishers; 2003.

[53] Cole JC, Rabin AS, Smith TL, Kaufman AS. Development and validation of a Rasch-derived CES-D short form. Psychol Assess. 2004;16(4):360-72.

[54] Cook KF, Teal CR, Bjorner JB, Cella D, Chang CH, Crane PK, et al. IRT health outcomes data analysis project: an overview and summary. Qual Life Res. 2007;16 Suppl 1:121-32.

[55] Coste J, Guillemin F, Pouchot J, Fermanian J. Methodological approaches to shortening composite measurement scales. J Clin Epidemiol. 1997;50(3):247-52.

[56] Cowen P, Anderson I, Fairburn CG. Neurochemical effects of dieting: Relevance to changes in eating and affective disorders. The biology of feast and famine: Relevance to eating disorders . San Diego, CA: Academic Press; 1992.

[57] Cronbach LJ. Coefficient alpha and the internal structure of test. Psychometrika. 1951;16:297-334.

[58] Cushnaghan J, Coggon D, Reading I, Croft P, Byng P, Cox K, et al. Long-term outcome following total hip arthroplasty: a controlled longitudinal study. Arthritis Rheum. 2007;57(8):1375-80.

[59] Davidson M, Keating JL, Eyres S. A low back-specific version of the SF-36 Physical Functioning scale. Spine. 2004;29(5):586-94.

[60] Davis AM, Badley EM, Beaton DE, Kopec J, Wright JG, Young NL, et al. Rasch analysis of the Western Ontario McMaster (WOMAC) Osteoarthritis Index: results from community and arthroplasty samples. J Clin Epidemiol. 2003;56(11):1076-83.

[61] Dawson J, Linsell L, Zondervan K, Rose P, Randall T, Carr A, et al. Epidemiology of hip and knee pain and its impact on overall health status in older adults. Rheumatology (Oxford). 2004;43(4):497-504.

[62] de Ayala RJ. The theory and practice of item response theory. New York: Guilford Press; 2009.

[63] de la Rie SM, Noordenbos G, Donker M, van FE. The patient's view on quality of life and eating disorders. Int J Eat Disord. 2006;40(1):13-20.

[64] de la Rie SM, Noordenbos G, van Furth EF. Quality of life and eating disorders. Qual Life Res. 2005;14(6):1511-22.

[65] de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine. Cambridge, UK: Cambridge University Press; 2011.

[66] DeMars C. Item response theory. New York: Oxford University Press; 2010.

[67] Devins GM, Dion R, Pelletier LG, Shapiro CM, Abbey S, Raiz LR, et al. Structure of lifestyle disruptions in chronic disease: a confirmatory factor analysis of the Illness Intrusiveness Ratings Scale. Med Care. 2001;39(10):1097-104.

[68] Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials. 1991;12(4 Suppl):142S-58S.

[69] Dillman DA. Mail and Telephone Surveys: The Total Design Method. New York: John Wiley & Sons; 1975.

[70] Doll HA, Petersen SE, Stewart-Brown SL. Eating disorders and emotional and physical well-being: associations between student self-reports of eating disorders and quality of life as measured by the SF-36. Qual Life Res. 2005;14(3):705-17.

[71] Donovan K, Sanson-Fisher RW, Redman S. Measuring quality of life in cancer patients. J Clin Oncol. 1989;7(7):959-68.

[72] Duncan PW, Lai SM, Bode RK, Perera S, DeRosa J. Stroke Impact Scale-16: A brief assessment of physical function. Neurology. 2003;60(2):291-6.

[73] Eignor DR. Standards for the development and use of tests: The Standards for Educational and Psychological Testing. European Journal of Psychological Assessment. 2001;17(3):157-63.

[74] Embretson S. A general latent trait model for response processes. Psychometrika. 1984;49(2):175-86.

[75] Embretson SE, Hershberger SL. The new rules of measurement: What every psychologist and educator should know. Mahwah, NJ: Lawrence Erlbaum Associates Publishers; 1999.

[76] Engel SG, Adair CE, Las Hayas C, Abraham S. Health-related quality of life and eating disorders: a review and update. Int J Eat Disord. 2009;42(2):179-87.

[77] Engel SG, Wittrock DA, Crosby RD, Wonderlich SA, Mitchell JE, Kolotkin RL. Development and psychometric validation of an eating disorder-specific health-related quality of life instrument. Int J Eat Disord. 2006;39(1):62-71.

[78] Epstein RS, Sherwood LM. From outcomes research to disease management: a guide for the perplexed. Ann Intern Med. 1996;124(9):832-7.

[79] Escobar A, Quintana JM, Bilbao A, Arostegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after total knee replacement. Osteoarthritis Cartilage. 2007b;15(3):273-80.

[80] Escobar A, Quintana JM, Bilbao A, Azkarate J, Guenaga JI, Arenaza JC, et al. Effect of patient characteristics on reported outcomes after total knee replacement. Rheumatology (Oxford). 2007a;46(1):112-9.

[81] Escobar A, Quintana JM, Bilbao A, Azkarate J, Guenaga JI. Validation of the Spanish version of the WOMAC questionnaire for patients with hip or knee osteoarthritis. Western Ontario and McMaster Universities Osteoarthritis Index. Clin Rheumatol. 2002;21(6):466-71.

[82] EuroQoL Group. EuroQol--a new facility for the measurement of health-related quality of life. Health Policy. 1990;16(3):199-208.

[83] Fairburn CG, Bohn K. Eating disorder NOS (EDNOS): an example of the troublesome "not otherwise specified" (NOS) category in DSM-IV. Behav Res Ther. 2005;43(6):691-701.

[84] Fairclough DL. Design and analysis of quality of life studies in clinical trials. Boca Raton, FL: Chapman & Hall/CRC; 2002.

[85] Fayers PM, Hand DJ. Factor analysis, causal indicators and quality of life. Qual Life Res. 1997;6(2):139-50.

[86] Fayers PM, Hand DJ, Bjordal K, Groenvold M. Causal indicators in quality of life research. Qual Life Res. 1997;6(5):393-406.

[87] Fayers PM, Machin D. Quality of life: assessment, analysis and interpretation. West Sussex, UK: Wiley; 2000.

[88] Fayers PM, Machin D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes (2nd ed.). London: John Wiley and Sons; 2007.

[89] Felson DT, Zhang Y. An update on the epidemiology of knee and hip osteoarthritis with a view to prevention. Arthritis Rheum. 1998;41(8):1343-55.

[90] Felson DT, Zhang Y, Hannan MT, Naimark A, Weissman BN, Aliabadi P, et al. The incidence and natural history of knee osteoarthritis in the elderly. The Framingham Osteoarthritis Study. Arthritis Rheum. 1995;38(10):1500-5.

[91] Fontaine KR, Bartlett SJ. Estimating health-related quality of life in obese individuals. Disease Management and Health Outcomes. 1998;3:61-70.

[92] Fontaine KR, Cheskin LJ, Barofsky I. Health-related quality of life in obese persons seeking treatment. J Fam Pract. 1996;43(3):265-70.

[93] Fontaine KR, Cheskin LJ, Barofsky I. Obesity and health-related quality of life. Obes Rev. 2001;2:219-29.

[94] Frankel S, Eachus J, Pearson N, Greenwood R, Chan P, Peters TJ, et al. Population requirement for primary hip-replacement surgery: a cross-sectional study. Lancet. 1999;353:1304-9.

[95] Gandek B, Ware JE, Jr. Methods for validating and norming translations of health status questionnaires: the IQOLA Project approach. International Quality of Life Assessment. J Clin Epidemiol. 1998;51(11):953-9.

[96] Gandek B, Ware JE, Aaronson NK, Apolone G, Bjorner JB, Brazier JE, et al. Cross-validation of item selection and scoring for the SF-12 Health Survey in

nine countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol. 1998;51(11):1171-8.

[97] Gao F, Luo N, Thumboo J, Fones C, Li SC, Cheung YB. Does the 12-item General Health Questionnaire contain multiple factors and do we need them? Health Qual Life Outcomes. 2004;2:63.

[98] Garner DM, Olmsted MP, Bohr Y, Garfinkel PE. The eating attitudes test: psychometric features and clinical correlates. Psychol Med. 1982;12(4):871-8.

[99] Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. Annals of Applied Statistics. 2008;2:1360-83.

[100] Goldsmith SB. The status of health status indicators. Health Serv Rep. 1972;87(3):212-20.

[101] Goldsmith SB. A reevaluation of health status indicators. Health Serv Rep. 1973;88(10):937-41.

[102] Gomez R. Parent rating of the ADHD items of the disruptive behaviour rating scale: Analyses of their IRT properties based on the generalized partial credit model. Personality and Individual Differences. 2008;45:181-6.

[103] Gomez R. Item response theory analyses of adult self-ratings of the ADHD symptoms in the Current Symptoms Scale. Assessment. 2011;18(4):476-86.

[104] Gonzalez N, Padierna A, Quintana JM, Arostegui I, Horcajo MJ. Quality of life in patients with eating disorders. Gac Sanit. 2001;15(1):18-24.

[105] Greenacre MJ. Theory and applications of correspondence analysis. London: Academic Press; 1984.

[106] Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol. 1993;46(12):1417-32.

[107] Gulliksen H. Theory of mental tests. New York: John Wiley; 1950.

[108] Guy W. Early clinical drug evaluation (ECDEU) assessment manual. Rockville, MD: National Institute Mental Health; 1976.

[109] Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Ann Intern Med. 1993;118(8):622-9.

[110] Hair JF, Tatham RL, Anderson RE, Black W. Multivariate data analysis (5th ed.). Upper Saddle River, NJ: Prentice Hall; 1998.

[111] Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. California: Sage Publications, Inc; 1991.

[112] Harrington D. Confirmatory factor analysis. New York: Oxford University Press, Inc.; 2009.

[113] Hatcher L. A step-by step approach to using the SAS System for Factor Analysis and Structural Equation Modeling. Cary, NC: SAS Institute Inc.; 1994.

[114] Hawker G, Melfi C, Paul J, Green R, Bombardier C. Comparison of a generic (SF-36) and a disease specific (WOMAC) (Western Ontario and McMaster Universities Osteoarthritis Index) instrument in the measurement of outcomes after knee replacement surgery. J Rheumatol. 1995;22(6):1193-6.

[115] Hawker G, Wright J, Coyte P, Paul J, Dittus R, Croxford R, et al. Health-related quality of life after knee replacement. J Bone Joint Surg Am. 1998;80(2):163-73.

[116] Hay PJ, Mond J. How to 'count the cost' and measure burden? A review of health-related quality of life in people with eating disorders. Journal of Mental Health. 2005;14(6):539-52.

[117] Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000;38(9 Suppl):II28-II42.

[118] Hillege S, Beale B, McMaster R. Impact of eating disorders on family life: individual parents' stories. J Clin Nurs. 2006;15(8):1016-22.

[119] Hoek HW. Incidence, prevalence and mortality of anorexia nervosa and other eating disorders. Curr Opin Psychiatry. 2006;19(4):389-94.

[120] Hoek HW, van Hoeken D. Review of the prevalence and incidence of eating disorders. Int J Eat Disord. 2003;34(4):383-96.

[121] Hsu LKG. Outcome of bulimia nervosa. In: Brownell KD, Fairburn CG, editors. Eating disorders and obesity: A comprehensive handbook. New York: Guildford Press; 1995. p. 238-44.

[122] Hu Lt, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling. 1999;6(1):1-55.

[123] Huang IC, Speight J. Psychometric Methods for Patient-reported Outcomes Assessment. In: Esposito D, Migliaccio-Walle K, Molsen E, editors. Reliability and Validity of Data Sources for Outcomes Research and Disease and Health Management Programs. Lawrenceville, NJ: International Society for Pharmaeconomics and Outcomes Research (ISPOR); 2013.

[124] Hudson JI, Hiripi E, Pope HG, Jr., Kessler RC. The prevalence and correlates of eating disorders in the National Comorbidity Survey Replication. Biol Psychiatry. 2007;61(3):348-58.

[125] Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol. 2000;53(5):459-68.

[126] Jenkins PE, Hoste RR, Meyer C, Blissett JM. Eating disorders and quality of life: a review of the literature. Clin Psychol Rev. 2011;31(1):113-21.

[127] Juniper EF, Guyatt GH, Streiner DL, King DR. Clinical impact versus factor analysis for quality of life questionnaire construction. J Clin Epidemiol. 1997;50(3):233-8.

[128] Kalantar JS, Talley NJ. The effects of lottery incentive and length of questionnaire on health survey response rates: a randomized study. J Clin Epidemiol. 1999;52(11):1117-22.

[129] Kaplan RM. The significance of quality of life in health care. Qual Life Res. 2003;12 Suppl 1:3-16.

[130] Karlsson J, Taft C, Sjostrom L, Torgerson JS, Sullivan M. Psychosocial functioning in the obese before and after weight reduction: construct validity and responsiveness of the Obesity-related Problems scale. Int J Obes Relat Metab Disord. 2003;27(5):617-30.

[131] Kaufman S. The emerging role of health-related quality of life: data in clinical research, part 2. Clin Res. 2001;1:38-43.

[132] Kaukua J, Pekkarinen T, Sane T, Mustajoki P. Health-related quality of life in WHO class II-III obese men losing weight with very-low-energy diet and behaviour modification: a randomised clinical trial. Int J Obes Relat Metab Disord. 2002;26(4):487-95.

[133] Keilen M, Treasure T, Schmidt U, Treasure J. Quality of life measurements in eating disorders, angina, and transplant candidates: are they comparable? J R Soc Med. 1994;87(8):441-4.

[134] Keller SD, Ware JE, Jr., 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. International Quality of Life Assessment. J Clin Epidemiol. 1998;51(11):1179-88.

[135] Kim JO, Mueller CW. Factor analysis: Statistical methods and practical issues. Beverly Hills, CA: Sage; 1978a.

[136] Kim JO, Mueller CW. Introduction to factor analysis: What it is and how to do it. Beverly Hills, CA: Sage; 1978b.

[137] Kline RB. Principles and Practice of Structural Equation Modelling (2nd ed.). New York: Guilford; 2005.

[138] Knutson K, Lewold S, Robertsson O, Lidgren L. The Swedish knee arthroplasty register. A nation-wide study of 30,003 knees 1976-1992. Acta Orthop Scand. 1994;65(4):375-86.

[139] Koller M, Lorenz W. Quality of life: a deconstruction for clinicians. J R Soc Med. 2002;95(10):481-8.

[140] Kolotkin RL, Crosby RD, Kosloski KD, Williams GR. Development of a brief measure to assess quality of life in obesity. Obes Res. 2001;9(2):102-11.

[141] Kolotkin RL, Head S, Brookhart A. Construct validity of the Impact of Weight on Quality of Life Questionnaire. Obes Res. 1997;5(5):434-41.

[142] Kolotkin RL, Head S, Hamilton M, Tse CK. Assessing Impact of Weight on Quality of Life. Obes Res. 1995;3(1):49-56.

[143] Kushner RF, Foster GD. Obesity and quality of life. Nutrition. 2000;16(10):947-52.

[144] Larsson U, Karlsson J, Sullivan M. Impact of overweight and obesity on health-related quality of life--a Swedish population study. Int J Obes Relat Metab Disord. 2002;26(3):417-24.

[145] Las Hayas C, Quintana JM, Padierna A, Bilbao A, Muñoz P, Madrazo A, et al. The new questionnaire health-related quality of life for eating disorders showed good validity and reliability. J Clin Epidemiol. 2006;59(2):192-200.

[146] Las Hayas C, Quintana JM, Padierna JA, Bilbao A, Muñoz P. Use of Rasch methodology to develop a short version of the health related quality of life for eating disorders questionnaire: a prospective study. Health Qual Life Outcomes. 2010;8:29.

[147] Las Hayas C, Quintana JM, Padierna JA, Bilbao A, Muñoz P, Cook FE. Health-Related Quality of Life for Eating Disorders questionnaire version-2 was responsive 1-year after initial assessment. J Clin Epidemiol. 2007;60(8):825-33.

[148] Le Pen C, Levy E, Loos F, Banzet MN, Basdevant A. "Specific" scale compared with "generic" scale: a double measurement of the quality of life in a French community sample of obese subjects. J Epidemiol Community Health. 1998;52(7):445-50.

[149] Lee SY, Poon WY, Bentler PM. A two-stage estimation of structural equation models with continuous and polytomous variables. Br J Math Stat Psychol. 1995;48 (Pt 2):339-58.

[150] Lee SY, Song XY. Evaluation of the Bayesian and Maximum Likelihood Approaches in Analyzing Structural Equation Models with Small Sample Sizes. Multivariate Behavioral Research. 2004;39(4):653-86.

[151] Lee SY, Song XY, Cai JH. A Bayesian approach for nonlinear structural equation models with dichotomous variables using logit and probit links. Structural Equation Modeling. 2010;17(2):280-302.

[152] Lewis KM, Lambert MC. Measuring social change preferences in African American adolescents: development of the Measure of Social Change for Adolescents (MOSC-A). Assessment. 2006;13(4):406-16.

[153] Linacre J. Sample size and item calibration stability. Rasch Meas Trans. 1994;7:328.

[154] Linacre J. Investigating rating scale category utility. J Outcome Meas. 1999;3(2):103-22.

[155] Linacre J. A user's guide to Winsteps / Ministeps Rasch-Model Programs. Chicago, IL: MESA Press; 2005.

[156] Linacre J. A user's guide to WINSTEPS. Chicago, IL: MESA Press; 2009.

[157] Long JS. Confirmatory Factor Analysis: A Preface to LISREL. Beverly Hills: Sage University paper Series on Quantitative Application in the Social Sciences; 1983.

[158] Lord FM, Novick MR. Statistical Theory of Mental Test Scores. Reading, MA: Addison-Wesley; 1968.

[159] MacCallum RC, Widaman KF, Zhang S, Hong S. Sample size in factor analysis. Psychological Methods. 1999;4(1):84-99.

[160] Mahon JL, Bourne RB, Rorabeck CH, Feeny DH, Stitt L, Webster-Bogaert S. Health-related quality of life and mobility of patients awaiting elective total hip arthroplasty: a prospective study. CMAJ. 2002;167(10):1115-21.

[161] Maïano C, Morin AJ, Lanfranchi MC, Therme P. The Eating Attitudes Test-26 revisited using exploratory structural equation modeling. J Abnorm Child Psychol. 2013;41(5):775-88.

[162] Mannucci E, Ricca V, Barciulli E, Di BM, Travaglini R, Cabras PL, et al. Quality of life and overweight: the obesity related well-being (Orwell 97) questionnaire. Addict Behav. 1999;24(3):345-57.

[163] Mar J, Begiristain JM, Arrazola A. Cost-effectiveness analysis of thrombolytic treatment for stroke. Cerebrovasc Dis. 2005a;20(3):193-200.

[164] Mar J, Rivero-Arias O, Duran-Cantolla J, Alonso-Alvarez ML, Gaminde I, de la Torre-Muñecas G. Effect on quality of life of nCPAP treatment in patients with obstructive sleep apnea. Med Clin (Barc). 2005b;125(16):611-5.

[165] Marquis P, Chassany O, Abetz L. A comprehensive strategy for the interpretation of quality-of-life data based on existing methods. Value Health. 2004;7(1):93-104.

[166] Masters GN. A Rasch model for partial credit scoring. Psychometrika. 1982;47(2):149-74.

[167] Mathias SD, Williamson CL, Colwell HH, Cisternas MG, Pasta DJ, Stolshek BS, et al. Assessing health-related quality-of-life and health state preference in persons with obesity: a validation study. Qual Life Res. 1997;6(4):311-22.

[168] Maydeu-Olivares A. Further Empirical Results on Parametric Versus Non-Parametric IRT Modeling of Likert-Type Personality Data. Multivariate Behavioral Research. 2005;40(2):261-79.

[169] McConnell S, Kolopack P, Davis AM. The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC): a review of its utility and measurement properties. Arthritis Rheum. 2001;45(5):453-61.

[170] McDonald RP. Future directions for item response theory. International Journal of Educational Research. 1989;13(2):205-20.

[171] Miller KK, Grinspoon SK, Ciampa J, Hier J, Herzog D, Klibanski A. Medical findings in outpatients with anorexia nervosa. Arch Intern Med. 2005 14;165(5):561-6.

[172] Miller PM. Redefining success in eating disorders. Addict Behav. 1996;21(6):745-54.

[173] Mitchison D, Hay P, Slewa-Younan S, Mond J. Time trends in population prevalence of eating disorder behaviors and their relationship to quality of life. PLoS One. 2012;7(11):e48450.

[174] Mond JM, Hay PJ, Rodgers B, Owen C, Beumont PJ. Assessing quality of life in eating disorder patients. Qual Life Res. 2005;14(1):171-8.

[175] Moran LA, Guyatt GH, Norman GR. Establishing the minimal number of items for a responsive, valid, health-related quality of life instrument. J Clin Epidemiol. 2001;54(6):571-9.

[176] Morrissey C, Cooke D, Michie C, Hollin C, Hogue T, Lindsay WR, et al. Structural, item, and test generalizability of the psychopathy checklist--revised to offenders with intellectual disabilities. Assessment. 2010;17(1):16-29.

[177] Mulaik SA, James LR, Van Alstine J, Bennett N, Lind S, Stilwell CD. Evaluation of goodness-of-fit indices for structural equation models. Psychological Bulletin. 1989;105(3):430-45.

[178] Muraki E. A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement. 1992;16(2):159-76.

[179] Muthén LK, Muthén BO. How to use a Monte Carlo study to decide on sample size and determine power. Structural Equation Modeling. 2002;9(4):599-620.

[180] Muthén LK, Muthén BO. Mplus User's Guide (6th ed.). Los Angeles, CA: Muthén & Muthén; 2010.

[181] Niero M, Martin M, Finger T, Lucas R, Mear I, Wild D, et al. A new approach to multicultural item generation in the development of two obesity-specific measures: the Obesity and Weight Loss Quality of Life (OWLQOL) questionnaire and the Weight-Related Symptom Measure (WRSM). Clin Ther. 2002;24(4):690-700.

[182] Nijsten T, Unaeze J, Stern RS. Refinement and reduction of the Impact of Psoriasis Questionnaire: classical test theory vs. Rasch analysis. Br J Dermatol. 2006;154(4):692-700.

[183] Nunnally JC. Psychometric theory (2nd ed.). New York: McGraw-Hill; 1978.

[184] Nunnally JC, Bernstein IH. Psychometric theory (3rd ed.). New York: McGraw-Hill; 1994.

[185] Núñez M, Lozano L, Núñez E, Segur JM, Sastre S, Macule F, et al. Total knee replacement and health-related quality of life: factors influencing long-term outcomes. Arthritis Rheum. 2009;61(8):1062-9.

[186] Oria HE, Moorehead MK. Bariatric analysis and reporting outcome system (BAROS). Obes Surg. 1998;8(5):487-99.

[187] Padierna A, Quintana JM, Arostegui I, Gonzalez N, Horcajo MJ. The health-related quality of life in eating disorders. Qual Life Res. 2000;9(6):667-74.

[188] Padierna A, Quintana JM, Arostegui I, Gonzalez N, Horcajo MJ. Changes in health related quality of life among patients treated for eating disorders. Qual Life Res. 2002;11(6):545-52.

[189] Padierna JA, Ecenarro R, Horcajo MJ, Rebolledo I. La recuperación de un trastorno de la conducta alimentaria. In: ACABE, editor. Surgiendo del remolino. La recuperación de un trastorno de la conducta alimentaria. Bizkaia: Berekintza; 2006. p. 17-27.

[190] Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. Med Care. 1989;27(3 Suppl):S217-S232.

[191] Polivy J. Psychological consequences of food restriction. J Am Diet Assoc. 1996;96(6):589-92.

[192] Prieto G, Delgado AR. Rasch-modelling: a test. Psicothema. 2003;15(1):94-100.

[193] Prieto L, Alonso J, Lamarca R. Classical Test Theory versus Rasch analysis for quality of life questionnaire reduction. Health Qual Life Outcomes. 2003;1:27.

[194] Quintana JM, Arostegui I, Azkarate J, Goenaga JI, Elexpe X, Letona J, et al. Evaluation of explicit criteria for total hip joint replacement. J Clin Epidemiol. 2000;53(12):1200-8.

[195] Quintana JM, Escobar A, Arostegui I, Bilbao A, Armendariz P, Lafuente I, et al. Prevalence of symptoms of knee or hip joints in older adults from the general population. Aging Clin Exp Res. 2008a;20(4):329-36.

[196] Quintana JM, Arostegui I, Escobar A, Azkarate J, Goenaga JI, Lafuente I. Prevalence of knee and hip osteoarthritis and the appropriateness of joint replacement in an older population. Arch Intern Med. 2008b;168(14):1576-84.

[197] Quintana JM, Escobar A, Arostegui I, Bilbao A, Azkarate J, Goenaga JI, et al. Health-related quality of life and appropriateness of knee or hip joint replacement. Arch Intern Med. 2006;166(2):220-6.

[198] Quintana JM, Escobar A, Bilbao A, Arostegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after hip joint replacement. Osteoarthritis Cartilage. 2005;13(12):1076-83.

[199] Raftery AE. Bayesian Model Selection in Social Research. In: Raftery A, editor. Sociological Methodology. Cambridge, MA: Blackwell; 1995. p. 111-63.

[200] Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research; 1960.

[201] Rat AC, Guillemin F, Osnowycz G, Delagoutte JP, Cuny C, Mainard D, et al. Total hip or knee replacement for osteoarthritis: mid- and long-term quality of life. Arthritis Care Res. 2010;62(1):54-62.

[202] Raykov T, Marcoulides GA. On multilevel model reliability estimation from the perspective of structural equation modeling. Structural Equation Modeling. 2006;13(1):130-41.

[203] Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care. 2007;45(5 Suppl 1):S22-S31.

[204] Revicki DA, Osoba D, Fairclough D, Barofsky I, Berzon R, Leidy NK, et al. Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. Qual Life Res. 2000;9(8):887-900.

[205] Rodgers J, Martin CR, Morse RC, Kendell K, Verrill M. An investigation into the psychometric properties of the Hospital Anxiety and Depression Scale in patients with breast cancer. Health Qual Life Outcomes. 2005;3:41.

[206] Rorty M, Yager J, Buckwalter JG, Rossotto E. Social support, social adjustment, and recovery status in bulimia nervosa. Int J Eat Disord. 1999;26(1):1-12.

[207] Rothenfluh DA, Reedwisch D, Muller U, Ganz R, Tennant A, Leunig M. Construct validity of a 12-item WOMAC for assessment of femoro-acetabular impingement and osteoarthritis of the hip. Osteoarthritis Cartilage. 2008;16(9):1032-8.

[208] Ryser L, Wright BD, Aeschlimann A, Mariacher-Gehler S, Stucki G. A new look at the Western Ontario and McMaster Universities Osteoarthritis Index using Rasch analysis. Arthritis Care Res. 1999;12(5):331-5.

[209] Samejima F. Estimation of latent trait ability using a response pattern of graded scores. Psychometrika Monograph Supplement No. 17. Richmond, VA: Psychometric Society; 1969.

[210] Samejima F. Normal ogive model on the continuous response level in the multidimensional latent space. Psychometrika. 1974;39(1):111-21.

[211] Sanz J. Value and quantification of quality of life in Medicine. Med Clin (Barc). 1991;96(2):66-9.

[212] SAS QC. SAS/QC User's Guide, Version 8, Volumes 1, 2, and 3. Cary, NC: SAS Institute Inc.; 1999.

[213] Shevlin M, Miles JNV. Effects of sample size, model specification and factor loadings on the GFI in confirmatory factor analysis. Personality and Individual Differences. 1998;25(1):85-90.

[214] Sijtsma K, Molenaar IW. Introduction to no-parametric item response theory (Vol. 5). Thousand Oaks, Calif, USA: Sage; 2002.

[215] Smith AB, Wright EP, Rush R, Stark DP, Velikova G, Selby PJ. Rasch analysis of the dimensional structure of the Hospital Anxiety and Depression Scale. Psychooncology. 2006;15(9):817-27.

[216] Sociedad Española de Reumatología. Estudio EPISER: prevalencia e impacto de las enfermedades reumáticas en la población adulta española. Madrid: EPISER; 2001.

[217] Spitzer RL, Kroenke K, Linzer M, Hahn SR, Williams JB, deGruy FVI, et al. Health-related quality of life in primary care patients with mental disorders. Results from the PRIME-MD 1000 Study. JAMA. 1995;274(19):1511-7.

[218] Staquet MJ, Hays RD, Fayers PM. Quality of life assessment in clinical trials. Oxford: Oxford Press; 1998.

[219] Steinhausen HC. The course and outcome of anorexia nervosa. In: Brownell KD, Fairburn CG, editors. Eating disorders and obesity. A comprehensive handbook. New York: Guildford Press; 1995. p. 234-7.

[220] Steinhausen HC. The outcome of anorexia nervosa in the 20th century. Am J Psychiatry. 2002;159(8):1284-93.

[221] Stice E. Review of the evidence for a sociocultural model of bulimia nervosa and an exploration of the mechanisms of action. Clinical Psychology Review. 1994;14(7):633-61.

[222] Streiner DL, Norman GR. Health measurement scales: A practical guide to their development and use. United States: Oxford University Press; 1989.

[223] Streiner DL, Norman GR. Health measurement scales: A practical guide to their development and use (3rd ed.). Oxford: Oxford University Press; 1995.

[224] Strober M, Freeman R, Morrell W. The long-term course of severe anorexia nervosa in adolescents: survival analysis of recovery, relapse, and outcome predictors over 10-15 years in a prospective study. Int J Eat Disord. 1997;22(4):339-60.

[225] Sullivan M, Karlsson J, Sjostrom L, Backman L, Bengtsson C, Bouchard C, et al. Swedish obese subjects (SOS)--an intervention study of obesity. Baseline evaluation of health and psychosocial functioning in the first 1743 subjects examined. Int J Obes Relat Metab Disord. 1993;17(9):503-12.

[226] Sun Y, Sturmer T, Gunther KP, Brenner H. Reliability and validity of clinical outcome measurements of osteoarthritis of the hip and knee--a review of the literature. Clin Rheumatol. 1997;16(2):185-98.

[227] Tang WK, Wong E, Chiu HF, Ungvari GS. Rasch analysis of the scoring scheme of the HADS Depression subscale in Chinese stroke patients. Psychiatry Res. 2007;150(1):97-103.

[228] Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value Health. 2004;7 Suppl 1:S22-S26.

[229] Terwee CB, Dekker FW, Mourits MP, Gerding MN, Baldeschi L, Kalmann R, et al. Interpretation and validity of changes in scores on the Graves' ophthalmopathy quality of life questionnaire (GO-QOL) after different treatments. Clin Endocrinol (Oxf). 2001;54(3):391-8.

[230] Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med. 2003;35(3):105-15.

[231] Testa MA. Interpretation of quality-of-life outcomes: issues that affect magnitude and meaning. Med Care. 2000;38(9 Suppl):II166-II174.

[232] Tirico PP, Stefano SC, Blay SL. Validity studies of quality of life instruments for eating disorders: systematic review of the literature. J Nerv Ment Dis. 2010;198(12):854-9.

[233] Tubach F, Baron G, Falissard B, Logeart I, Dougados M, Bellamy N, et al. Using patients' and rheumatologists' opinions to specify a short form of the WOMAC function subscale. Ann Rheum Dis. 2005;64(1):75-9.

[234] van Hoeken D, Lucas AR, Hoek HW. Epidemiology. In: Hoek HW, Trasure JL, Katzman MA, editors. Neurobiology in the treatment of eating disorders. Chichester: John Wiley & Sons; 1998. p. 97-126.

[235] Vazquez JA, Gaztambide S, Soto-Pedre E. 10-year prospective study on the incidence and risk factors for type 2 diabetes mellitus. Med Clin (Barc). 2000;115(14):534-9.

[236] Velicer WF, Fava JL. Affects of variable and subject sampling on factor pattern recovery. Psychological Methods. 1998;3(2):231-51.

[237] Wadden TA, Phelan S. Assessment of quality of life in obese individuals. Obes Res. 2002;10 Suppl 1:50S-7S.

[238] Wade TD, Wilksch SM, Lee C. A longitudinal investigation of the impact of disordered eating on young women's quality of life. Health Psychol. 2012;31(3):352-9.

[239] Ware J, Jr., Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care. 1996;34(3):220-33.

[240] Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992;30(6):473-83.

[241] Whitehouse SL, Crawford RW, Learmonth ID. Validation for the reduced Western Ontario and McMaster Universities Osteoarthritis Index function scale. J Orthop Surg (Hong Kong). 2008;16(1):50-3.

[242] Whitehouse SL, Lingard EA, Katz JN, Learmonth ID. Development and testing of a reduced WOMAC function scale. J Bone Joint Surg Br. 2003;85(5):706-11.

[243] Wiebe S, Guyatt G, Weaver B, Matijevic S, Sidwell C. Comparative responsiveness of generic and specific quality-of-life instruments. J Clin Epidemiol. 2003;56(1):52-60.

[244] Wolfe F, Kong SX. Rasch analysis of the Western Ontario MacMaster questionnaire (WOMAC) in 2205 patients with osteoarthritis, rheumatoid arthritis, and fibromyalgia. Ann Rheum Dis. 1999;58(9):563-8.

[245] World Health Organization. In: World Health Organization. Handbook of Basic Documents (5th ed.). Ginebra: Palais des Nations; 1952. p. 3-20.

[246] Wright B, Stone M. Best test design: Rasch measurement. Chicago: MESA Press; 1979.

[247] Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol. 1999;52(9):861-73.

[248] Yang KG, Raijmakers NJ, Verbout AJ, Dhert WJ, Saris DB. Validation of the short-form WOMAC function scale for the evaluation of osteoarthritis of the knee. J Bone Joint Surg Br. 2007;89(1):50-6.

Anexos

Anexo I. Cuestionario Obesity-related Problems scale (OP)

A continuación se presenta la versión española del cuestionario Obesity-related Problems scale (OP) diseñado para medir el impacto de la obesidad mórbida en el funcionamiento psicosocial.

Escala de problemas relacionados con la obesidad

¿La obesidad le molesta cuando realiza las siguientes actividades?

Marque la alternativa que mejor refleje su situación

	Me molesta mucho	Me molesta bastante	Me molesta poco	No me molesta en absoluto
1. Reuniones particulares en mi propia casa	1☐	2☐	3☐	4☐
2. Reuniones en cada de amigos o familiares	1☐	2☐	3☐	4☐

	Me molesta mucho	Me molesta bastante	Me molesta poco	No me molesta en absoluto
3. Cuando voy a un restaurantes	1☐	2☐	3☐	4☐
4. Cuando asisto a actividades sociales, cursos,…	1☐	2☐	3☐	4☐
5. Cuando voy de vacaciones fuera de casa	1☐	2☐	3☐	4☐
6. Al probarse y comprar ropa	1☐	2☐	3☐	4☐
7. Al bañarse en lugares públicos (playa, piscina, etc.)	1☐	2☐	3☐	4☐
8. Relaciones íntimas con su pareja	1☐	2☐	3☐	4☐
Si no tiene pareja, marque esta casilla ☐				

Anexo II. Cuestionario Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)

A continuación se presenta la versión española del cuestionario Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) diseñado para medir sintomatología y función en pacientes con artrosis de extremidad inferior.

CUESTIONARIO WOMAC

Las preguntas que siguen se refieren al dolor, rigidez y dificultad en la realización de diferentes actividades que usted pudiera tener **debido a sus problemas de cadera**.

Para cada una de las siguientes preguntas, elija sólo una respuesta y márquela poniendo una X sobre el cuadrado. Conteste por favor todas las preguntas. Si no está seguro/a de cómo responder a una pregunta, por favor conteste lo que le parezca más cierto.

1. Las siguientes preguntas tratan sobre la **intensidad del dolor** que ha tenido durante el **último mes** en la/s **cadera/s afectada/s**. Si no realiza alguna de las actividades, contéstela pensando como cree usted que podría realizarla.

Pregunta: ¿Cuánto dolor tiene?

a) Al **andar** por un **terreno llano**

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

b) Al **subir o bajar escaleras**

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

c) Por la noche en la cama

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

d) Al estar **sentado o tumbado**

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

e) Al estar **de pie**

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

2. Las siguientes preguntas tratan sobre la intensidad de la **rigidez** articular (se refiere a la dificultad para mover la cadera, no al dolor) que usted ha tenido durante **el último mes** en la/s **cadera/s afectada/s**.

a) ¿Cuánta **rigidez** nota **después de despertarse por la mañana**?

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

b) ¿Cuánta **rigidez** nota durante el **resto del día después de estar sentado, tumbado o descansado**?

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

3. Las siguientes preguntas se refieren a la **dificultad** que usted ha tenido **para hacer ciertas cosas durante el último mes**. Si no realiza alguna de las actividades, contéstela pensando cómo cree usted que podría realizarla.

Pregunta: ¿Que grado de dificultad tiene al…?

a) Bajar escaleras

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

b) Subir las escaleras

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

c) Levantarse después de estar sentado

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

d) Estar de pie

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

e) Agacharse para coger algo del suelo

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

f) Andar por un terreno llano

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

g) Entrar y salir de un coche

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

h) Ir de compras

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

i) Ponerse los calcetines / medias

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

j) Levantarse de la cama

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

k) Quitarse los calcetines / medias

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

l) Estar tumbado en la cama

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

m) Entrar y salir de la ducha / bañera

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

n) Estar sentado

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

o) Sentarse y levantarse del retrete, inodoro

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

p) Hacer tareas o actividades pesadas

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

q) Hacer tareas o actividades sencillas

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

Anexo III. Cuestionario Health Related Quality of Life for Eating Disorders version 2 (HeRQoLEDv2)

A continuación se presenta el cuestionario Health Related Quality of Life for Eating Disorders version 2 (HeRQoLEDv2) diseñado para medir la calidad de vida relacionada con la salud en pacientes con un trastornod e la conducta de la alimentación.

En el cuestionario que tienes en tus manos encontrarás algunas preguntas sobre tu calidad de vida, tu alimentación y diferentes aspectos de tu vida diaria.

Por favor, es importante que contestes a todas las preguntas marcando con una X la alternativa que elijas. No hay respuestas correctas o incorrectas. Si no estás segura de cómo responder a una pregunta, da la respuesta que mejor describa tu situación. Completar todas las preguntas te supondrá aproximadamente 20 minutos.

Todas las respuestas que nos des serán tratadas con la más absoluta confidencialidad.

EJEMPLO

En cada pregunta, por favor, marca con una X la respuesta que elijas.

En las últimas 4 semanas, ¿has tenido dificultades para dormir?

Nunca	Casi nunca	Algunas veces	Bastantes veces	Casi siempre	Siempre
☒	☐	☐	☐	☐	☐

Si te equivocas o cambias de opinión, rodea con un círculo la respuesta incorrecta y vuelve a marcar con una X tu respuesta definitiva.

En las últimas 4 semanas, ¿has tenido dificultades para dormir?

Nunca	Casi nunca	Algunas veces	Bastantes veces	Casi siempre	Siempre
⊗	☒	☐	☐	☐	☐

Estas primeras preguntas quieren conocer si has tenido alguno de los siguientes síntomas.

Durante las <u>últimas 4 semanas</u>, ¿has tenido alguno de los siguientes síntomas o molestias?

	Nunca (0)	Casi nunca (1)	Algunas veces (2)	Bastantes veces (3)	Constante mente (4)
1. Palpitaciones	☐	☐	☐	☐	☐
2. Mareos	☐	☐	☐	☐	☐
3. Dificultad para respirar	☐	☐	☐	☐	☐
4. Dolor muscular	☐	☐	☐	☐	☐
5. Digestiones pesadas	☐	☐	☐	☐	☐
6. Estreñimiento	☐	☐	☐	☐	☐
7. Mayor sensación de frío o soportar peor el frío	☐	☐	☐	☐	☐
8. Cansancio físico	☐	☐	☐	☐	☐

9. Durante los <u>últimos 3 meses</u> y en relación a lo que tú consideras normal, ¿has tenido problemas porque se te caía el pelo excesivamente?

☐ 0 Nunca
☐ 1 Casi Nunca
☐ 2 Algunas veces
☐ 3 Bastantes veces
☐ 4 Constantemente

10 ®. Durante <u>los últimos 3 meses</u>, ¿has tenido problemas en los dientes y/o en la boca?

☐ 4 Sí, graves
☐ 3 Sí, moderados
☐ 2 Sí, leves
☐ 0 No, ninguno.

11. Durante los <u>últimos 3 meses</u>, ¿has tenido la regla?

☐ Sí, los tres meses

☐ Una o dos reglas durante los 3 últimos meses

☐ No he tenido regla en los 3 últimos meses

☐ Necesito tomar anticonceptivos (píldora) para tener la regla.

☐ No tengo la regla por embarazo, operación u otras circunstancias. Por favor, especifica la razón: _____

☐ Tomo anticonceptivos.

Las siguientes preguntas se refieren a diferentes comportamientos que pueden presentarse con relación a la alimentación.

Durante las <u>últimas 4 semanas</u>, ¿has intentado controlar o reducir tu peso (no engordar) y/o alcanzar una figura más delgada, de alguna de las siguientes maneras?

	Nunca (0)	1 vez al <u>mes</u> (1)	2-3 veces al <u>mes</u> (2)	1-2 veces a la <u>semana</u> (3)	3-6 veces a la <u>semana</u> (4)	Todos los <u>días</u> (5)
12. Estando algún día sin comer, aunque tuvieses hambre	☐	☐	☐	☐	☐	☐
13. Saltándote algunas comidas, aunque tuvieses hambre	☐	☐	☐	☐	☐	☐
14. Comiendo exclusivamente alimentos bajos en calorías	☐	☐	☐	☐	☐	☐
15. Evitando comer con otras personas	☐	☐	☐	☐	☐	☐
16. Comiendo cada día los mismos alimentos	☐	☐	☐	☐	☐	☐
17. Realizando ejercicio en exceso para "quemar calorías"	☐	☐	☐	☐	☐	☐

18. En los <u>últimos 3 meses</u>, ¿ha habido ocasiones que has comido lo que otras personas considerarían una gran cantidad de comida, en un corto espacio de tiempo junto con una sensación de no poder parar de comer?

* Un **<u>atracón</u>** se caracteriza por tener <u>fuertes impulsos de comer</u>, durante los cuales ingieres en un **<u>breve espacio de tiempo</u>** (por lo general, menos de dos horas**<u>) grandes cantidades de comida</u>** (que sería considerada también una gran cantidad por otras personas) <u>o</u> grandes cantidades <u>de alimentos que generalmente te prohibes</u>, sintiendo que eres **<u>incapaz de parar</u>** de comer.

☐ 0 Nunca
☐ 1 Menos de 1 vez a la semana
☐ 2 Alrededor de 2 veces a la semana
☐ 3 Casi todos los días
☐ 4 Todos los días, una vez al día
☐ 5 Todos los días, varias veces al día

19. ¿Con qué frecuencia <u>piensas</u> en darte un atracón?

☐ 0 Nunca
☐ 1 Menos de 1 vez a la semana
☐ 2 Alrededor de 2 veces a la semana
☐ 3 Casi todos los días
☐ 4 Todos los días, una vez al día
☐ 5 Todos los días, varias veces al día

20. ¿Te provocas el vómito, o empleas otros métodos (como laxantes, diuréticos etc.) para evitar un aumento de peso o para aliviar la sensación de malestar/culpa?

☐ 0 Nunca
☐ 1 Menos de 1 vez a la semana
☐ 2 Alrededor de 2 veces a la semana
☐ 3 Casi todos los días
☐ 4 Todos los días, una vez al día
☐ 5 Todos los días, varias veces al día

21. ¿Con qué frecuencia has <u>pensado</u> en vomitar o emplear otros métodos (laxantes, diuréticos u otros) para evitar un aumento de peso o para aliviar la sensación de malestar/ culpa?

☐ 0 Nunca
☐ 1 Menos de 1 vez a la semana
☐ 2 Alrededor de 2 veces a la semana
☐ 3 Casi todos los días
☐ 4 Todos los días, una vez al día
☐ 5 Todos los días, varias veces al día

22. ¿Sueles comer <u>en exceso</u> cuando estas disgustada o estresada?

- ☐ 0 Nunca
- ☐ 1 Raras veces
- ☐ 2 Pocas veces
- ☐ 3 Algunas veces
- ☐ 4 Muchas veces
- ☐ 5 Siempre

23. ¿Crees que tener que seguir alguno de los COMPORTAMIENTOS anteriores (desde la pregunta 12 a la 22), afecta negativamente a tu Calidad de Vida?

- ☐ Nada
- ☐ Un poco
- ☐ Algo
- ☐ Bastante
- ☐ Mucho

- ☐ No tengo ninguno de estos hábitos

En el siguiente bloque te presentamos algunas preguntas sobre la percepción que tienes de tu cuerpo.

<u>En las últimas 4 semanas…</u>

24. En general, ¿te has sentido gorda, a pesar de que otras personas (familiares, amigos, médicos, etc.) te dicen que no lo estás?

- ☐ 0 Nunca
- ☐ 1 Casi nunca
- ☐ 2 Algunas veces
- ☐ 3 Bastantes veces
- ☐ 4 Casi siempre
- ☐ 5 Siempre

25. ¿Has pensado que algunas partes de tu cuerpo, por ejemplo, cadera, cintura o muslos, son excesivamente grandes o anchas respecto al resto del cuerpo?

- ☐ 0 Nunca
- ☐ 1 Casi nunca
- ☐ 2 Algunas veces
- ☐ 3 Bastantes veces
- ☐ 4 Casi siempre
- ☐ 5 Siempre

26. ¿Te has sentido preocupada por tu peso?

- ☐ 0 Nunca
- ☐ 1 Casi nunca
- ☐ 2 Algunas veces
- ☐ 3 Bastantes veces
- ☐ 4 Casi siempre
- ☐ 5 Siempre

27. ¿Has estado preocupada por la posibilidad de ganar peso o engordar?

- ☐ 0 Nunca
- ☐ 1 Casi nunca
- ☐ 2 Algunas veces
- ☐ 3 Bastantes veces
- ☐ 4 Casi siempre
- ☐ 5 Siempre

28. ¿Has evitado situaciones en las que otros puedan ver tu cuerpo como, por ejemplo, en el gimnasio, en la piscina o en la playa?

- ☐ 0 Nunca
- ☐ 1 Casi nunca
- ☐ 2 Algunas veces
- ☐ 3 Bastantes veces
- ☐ 4 Casi siempre
- ☐ 5 Siempre

29. ¿Te has sentido incómoda viendo tu cuerpo en un espejo, reflejado en un cristal (en una ventana, en un escaparate), al desnudarte o al ducharte?

- ☐ 0 Nunca
- ☐ 1 Casi nunca
- ☐ 2 Algunas veces
- ☐ 3 Bastantes veces
- ☐ 4 Casi siempre
- ☐ 5 Siempre

30. El estar preocupada por tu peso o figura, ¿ha hecho que te quedaras en casa?

- ☐ 0 Nunca
- ☐ 1 Casi nunca
- ☐ 2 Algunas veces
- ☐ 3 Bastantes veces
- ☐ 4 Casi siempre
- ☐ 5 Siempre

31. ®¿Estás satisfecha con tu apariencia física?

- ☐ 0 Nunca
- ☐ 1 Casi nunca
- ☐ 2 Algunas veces
- ☐ 3 Bastantes veces
- ☐ 4 Casi siempre
- ☐ 5 Siempre

Las preguntas que siguen se refieren a cómo te has sentido y cómo te han ido las cosas durante las últimas 4 semanas. En cada pregunta responde lo que se parezca más a cómo te has sentido.

Durante las <u>últimas 4 semanas</u>, ¿cuánto tiempo...

	Nunca (0)	Casi nunca (1)	Algunas veces (2)	Bastantes veces (3)	Casi siempre (4)	Siempre (5)
32. has estado muy nerviosa?	☐	☐	☐	☐	☐	☐
33. te has sentido tan baja de moral que nada podía animarte?	☐	☐	☐	☐	☐	☐
34. ® te has sentido calmada y tranquila?	☐	☐	☐	☐	☐	☐
35. te has sentido desanimada y triste?	☐	☐	☐	☐	☐	☐
36. ® te has sentido feliz?	☐	☐	☐	☐	☐	☐
37. has tenido cambios muy bruscos de humor que te ha costado controlar?	☐	☐	☐	☐	☐	☐

38. En las últimas 4 semanas, ¿has tenido dificultades para dormir <u>sin medicación</u>?

- ☐ 0 Nunca
- ☐ 1 Casi nunca
- ☐ 2 Algunas veces
- ☐ 3 Bastantes veces
- ☐ 4 Casi siempre
- ☐ 5 Siempre

39. ®En las últimas 4 semanas, ¿sientes que has dormido y descansado lo suficiente como para estar despejada durante el resto del día?

- ☐ 5 Nunca
- ☐ 4 Casi nunca
- ☐ 3 Algunas veces
- ☐ 2 Bastantes veces
- ☐ 1 Casi siempre
- ☐ 0 Siempre

40. ¿Has sentido que no vales nada?

- ☐ 0 Nunca
- ☐ 1 Casi nunca
- ☐ 2 Algunas veces
- ☐ 3 Bastantes veces
- ☐ 4 Casi siempre
- ☐ 5 Siempre

En tus estudios, trabajo o actividades cotidianas y debido a algún PROBLEMA EMOCIONAL, (como estar triste, deprimida o nerviosa), durante las <u>últimas 4 semanas</u>…

	Nunca (0)	Casi nunca (1)	Algunas veces (2)	Bastantes veces (3)	Casi siempre (4)	Siempre (5)
41. ¿has tenido que esforzarte más y/o dedicar más tiempo de lo habitual por algún problema emocional?	☐	☐	☐	☐	☐	☐
42. ¿has conseguido o has rendido menos de lo que hubieras querido por algún problema emocional?	☐	☐	☐	☐	☐	☐
43. ¿te ha costado más mantener la atención por algún problema emocional?	☐	☐	☐	☐	☐	☐
44. ¿has tenido que dejar de hacer algunas tareas por algún problema emocional?	☐	☐	☐	☐	☐	☐

En tus estudios, trabajo o actividades cotidianas y debido a algún PROBLEMA FÍSICO (como estar cansada, no haber podido dormir bien, tener mareos…), durante las <u>últimas 4 semanas</u>…

	Nunca (0)	Casi nunca (1)	Algunas veces (2)	Bastantes veces (3)	Casi siempre (4)	Siempre (5)
45. ¿has tenido que esforzarte más y/o dedicar más tiempo de lo habitual por algún problema físico?	☐	☐	☐	☐	☐	☐
46. ¿has conseguido o has rendido menos de lo que hubieras querido por algún problema físico?	☐	☐	☐	☐	☐	☐
47. ¿te ha costado más mantener la atención por algún problema físico?	☐	☐	☐	☐	☐	☐
48. ¿has tenido que dejar de hacer algunas tareas por algún problema físico?	☐	☐	☐	☐	☐	☐

En este bloque te presentamos algunas afirmaciones sobre tu forma de ser para que indiques en qué medida te identificas con ellas.

<u>En los últimos 3 meses…</u>

49. ¿Crees que te ha faltado confianza en tus propias capacidades?

☐ 0 Nunca
☐ 1 Casi nunca
☐ 2 Algunas veces
☐ 3 Bastantes veces
☐ 4 Casi siempre
☐ 5 Siempre

50. ¿Has creído que tenías que hacer las cosas perfectamente o no hacerlas?

☐ 0 Nunca
☐ 1 Casi nunca
☐ 2 Algunas veces
☐ 3 Bastantes veces
☐ 4 Casi siempre
☐ 5 Siempre

51. ¿Te has fijado objetivos muy exigentes y si no los has conseguido te has sentido insatisfecha?

- ☐ 0 Nunca
- ☐ 1 Casi nunca
- ☐ 2 Algunas veces
- ☐ 3 Bastantes veces
- ☐ 4 Casi siempre
- ☐ 5 Siempre

52. ¿Has pensado que los demás esperan de ti resultados sobresalientes?

- ☐ 0 Nunca
- ☐ 1 Casi nunca
- ☐ 2 Algunas veces
- ☐ 3 Bastantes veces
- ☐ 4 Casi siempre
- ☐ 5 Siempre

Las siguientes preguntas se refieren a cómo son <u>EN LA ACTUALIDAD</u> tus relaciones sociales habituales con tu familia, amigos, pareja…

53. ¿En qué medida tus preocupaciones en torno a la alimentación afectan negativamente tu <u>vida social habitual</u> (relaciones con conocidos, compañeros de trabajo o estudios…)?

- ☐ 0 Nada
- ☐ 1 Un poco
- ☐ 2 Algo
- ☐ 3 Bastante
- ☐ 4 Mucho

54. ¿En qué medida tus preocupaciones en torno a la alimentación afectan negativamente a tu <u>relación familiar</u> (hablar menos, más discusiones, menos confianza, etc.)?

- ☐ 0 Nada
- ☐ 1 Un poco
- ☐ 2 Algo
- ☐ 3 Bastante
- ☐ 4 Mucho

55. ¿En qué medida tus preocupaciones en torno a la alimentación afectan negativamente a tus relaciones con amigos cercanos (exceptuando tu pareja)?

☐ 0 Nada
☐ 1 Un poco
☐ 2 Algo
☐ 3 Bastante
☐ 4 Mucho

56. ¿Crees que tu problema de alimentación afecta negativamente tu relación de pareja o la posibilidad de encontrarla?

☐ 0 Nada
☐ 1 Un poco
☐ 2 Algo
☐ 3 Bastante
☐ 4 Mucho

57. ¿Crees que tu interés hacia el sexo está alterado por tu problema de alimentación?

☐ 0 Nada
☐ 1 Un poco
☐ 2 Algo
☐ 3 Bastante
☐ 4 Mucho

58. Las siguientes frases se refieren a tu actitud hacia la alimentación. Por favor, marca la frase con la que más te identifiques:

☐ No tengo ni he tenido ningún problema con la alimentación ni me pasa ni me ha pasado nada que necesite cambiar

☐ Pienso que tengo algún problema con la alimentación pero no he intentado nada para solucionarlo

☐ Creo que tengo que hacer algo para solucionar este problema con la alimentación que me preocupa

☐ Estoy trabajando para cambiar mi problema de alimentación

☐ He conseguido solucionar mi problema con la alimentación totalmente o casi totalmente

Por último, por favor, conteste a estas preguntas que piden datos generales sobre ti. Esta información, al igual que la que nos ha facilitado a lo largo de toda la encuesta, es completamente confidencial.

Edad: ..... **años.**

Peso actual: ..... **Kg.**

Altura actual: ..... **cm.**

Sexo: ☐ Mujer ☐ Hombre

¿Cuál es tu <u>estado civil actual</u>?

- ☐ Soltera/o
- ☐ Casada/o o conviviendo con la pareja
- ☐ Viuda/o
- ☐ Separada/o o divorciada/o

¿Cuál es tu <u>nivel de estudios</u>? (señala el nivel de estudios que hayas alcanzado hasta el momento actual)

- ☐ Sin estudios
- ☐ Estudios primarios (ESO / hasta 8º de EGB)
- ☐ Estudios secundarios (LOGSE / FP, REM, BUP...)
- ☐ Estudios superiores (universitarios)
- ☐ He finalizado estudios superiores (universitarios)

¿Cuál es tu <u>situación laboral</u> actual? (elige la o las respuestas que sean apropiadas)

- ☐ Trabajo a tiempo completo
- ☐ Trabajo a tiempo parcial
- ☐ Desempleada
- ☐ Estudiante (ó estás matriculado en algún curso)
- ☐ Incapacidad laboral
- ☐ Ama de casa
- ☐ Baja laboral

Fecha en que rellenas el cuestionario: _____/_____/_____ (día/mes/año)

Si quieres hacer algún comentario sobre el cuestionario en general o sobre alguna pregunta en concreto, puedes utilizar este espacio.

Número de pregunta	Comentarios

MUCHAS GRACIAS POR TU COLABORACIÓN

Anexo IV. Versión reducida del cuestionario Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)

A continuación se presenta la versión reducida del cuestionario Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) diseñado para medir dolor y función en pacientes con artrosis de extremidad inferior.

CUESTIONARIO REDUCIDO WOMAC

Las preguntas que siguen se refieren al dolor y dificultad en la realización de diferentes actividades que usted pudiera tener **debido a sus problemas de cadera**.

Para cada una de las siguientes preguntas, elija sólo una respuesta y márquela poniendo una X sobre el cuadrado. Conteste por favor todas las preguntas. Si no está seguro/a de cómo responder a una pregunta, por favor conteste lo que le parezca más cierto.

1. Las siguientes preguntas tratan sobre la **intensidad del dolor** que ha tenido durante el **último mes** en la/s **cadera/s afectada/s**. Si no realiza alguna de las actividades, contéstela pensando como cree usted que podría realizarla.

Pregunta: ¿Cuánto dolor tiene?

a) Al **andar** por un **terreno llano**

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

b) Al **subir o bajar escaleras**

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

c) Al estar **sentado o tumbado**

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

2. Las siguientes preguntas se refieren a la **dificultad** que usted ha tenido **para hacer ciertas cosas durante el último mes**. Si no realiza alguna de las actividades, contéstela pensando cómo cree usted que podría realizarla.

Pregunta: ¿Que grado de dificultad tiene al…?

a) Bajar escaleras

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

b) Subir las escaleras

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

c) Levantarse después de estar sentado

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

d) Andar por un terreno llano

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

e) Entrar y salir de un coche

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

f) Ir de compras

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

g) Ponerse los calcetines / medias

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

h) Sentarse y levantarse del retrete, inodoro

Ninguno	Poco	Bastante	Mucho	Muchísimo
☐	☐	☐	☐	☐

Anexo V. Cuestionario Health Related Quality of Life for Eating Disorders–Short Version (HeRQoLED-S)

A continuación se presenta el cuestionario reducido Health Related Quality of Life for Eating Disorders–Short Version (HeRQoLED-S) diseñado para medir la calidad de vida relacionada con la salud en pacientes con un trastornod e la conducta de la alimentación.

MARCA CON UNA 'X' LA CASILLA QUE MÁS SE AJUSTE A TU SITUACIÓN

- ***Durante las últimas 4 semanas, ¿has intentado controlar o reducir tu peso (no engordar) y/o alcanzar una figura más delgada, de alguna de las siguientes maneras?***

❶ **Estando algún día <u>sin comer</u>, aunque tuvieses <u>hambre.</u>**

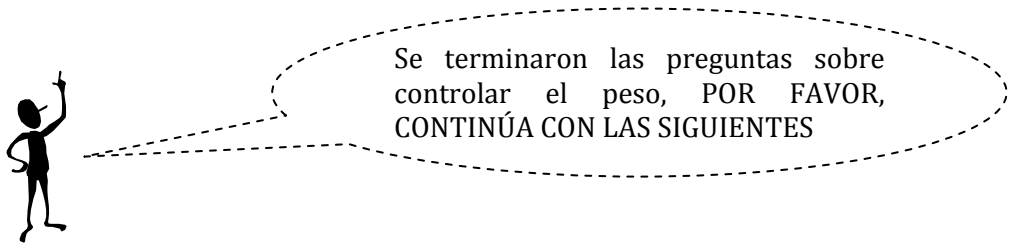
☐ Nunca ☐ Casi nunca ☐ Algunas veces ☐ Bastantes veces ☐ Casi siempre ☐ Siempre

❷ **<u>Saltándote</u> algunas <u>comidas</u>, aunque tuvieses <u>hambre.</u>**

☐ Nunca ☐ Casi nunca ☐ Algunas veces ☐ Bastantes veces ☐ Casi siempre ☐ Siempre

❸ **Evitando comer <u>con otras personas</u>.**

☐ Nunca ☐ Casi nunca ☐ Algunas veces ☐ Bastantes veces ☐ Casi siempre ☐ Siempre



- ***En las últimas 4 semanas…***

❹ **En general, ¿te has sentido gorda, <u>a pesar de que otras personas</u> (familiares, amigos, médicos, etc.) <u>te dicen que no lo estás</u>?**

☐ Nunca ☐ Casi nunca ☐ Algunas veces ☐ Bastantes veces ☐ Casi siempre ☐ Siempre

❺ **¿Has pensado que algunas <u>partes de tu cuerpo</u>, por ejemplo, cadera, cintura o muslos, son <u>excesivamente grandes o anchas</u> respecto al resto del cuerpo?**

□ Nunca □ Casi nunca □ Algunas veces □ Bastantes veces □ Casi siempre □ Siempre

❻ **¿Te has sentido <u>preocupada</u> por tu peso?**

□ Nunca □ Casi nunca □ Algunas veces □ Bastantes veces □ Casi siempre □ Siempre

❼ **¿Has estado <u>preocupada</u> por la posibilidad de ganar peso o <u>engordar</u>?**

□ Nunca □ Casi nunca □ Algunas veces □ Bastantes veces □ Casi siempre □ Siempre

❽ **¿Has <u>evitado situaciones</u> en las que <u>otros puedan ver tu cuerpo</u> como, por ejemplo, en el gimnasio, en la piscina o en la playa?**

□ Nunca □ Casi nunca □ Algunas veces □ Bastantes veces □ Casi siempre □ Siempre

❾ **¿En qué medida <u>tus preocupaciones en torno a la alimentación</u> afectan negativamente a tu <u>relación familiar</u> (hablar menos, más discusiones, menos confianza, etc.)?**

□ Nunca □ Casi nunca □ Algunas veces □ Bastantes veces □ Casi siempre □ Siempre

❿ **¿Crees que tu <u>problema de alimentación</u> afecta negativamente tu <u>relación de pareja</u> o la posibilidad de encontrarla?**

□ Nada □ Un poco □ Algo □ Bastante □ Mucho

SÓLO QUEDAN ESTAS 10 ÚLTIMAS PREGUNTAS:

- ***Durante las últimas 4 semanas, ¿cuánto tiempo…***

① **¿Te has <u>sentido feliz</u>?**

□ Nunca □ Casi nunca □ Algunas veces □ Bastantes veces □ Casi siempre □ Siempre

② **¿Has tenido <u>cambios muy bruscos de humor</u> que te ha costado controlar?**

□ Nunca □ Casi nunca □ Algunas veces □ Bastantes veces □ Casi siempre □ Siempre

③ **¿Has sentido que no vales nada?**

□ Nunca □ Casi nunca □ Algunas veces □ Bastantes veces □ Casi siempre □ Siempre

- ***En tus estudios, trabajo o actividades cotidianas y debido a algún <u>PROBLEMA EMOCIONAL</u>, (como estar triste, deprimida o nerviosa), durante las últimas 4 semanas…***

④ **¿Has tenido <u>que esforzarte más y/o dedicar más tiempo de lo habitual</u> por algún problema emocional?**

□ Nunca □ Casi nunca □ Algunas veces □ Bastantes veces □ Casi siempre □ Siempre

⑤ **¿Has <u>conseguido o has rendido menos</u> de lo que hubieras querido por algún problema emocional?**

□ Nunca □ Casi nunca □ Algunas veces □ Bastantes veces □ Casi siempre □ Siempre

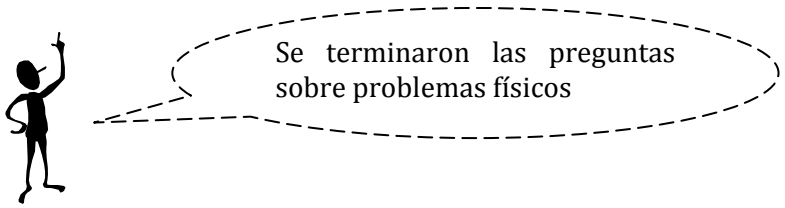
- ***En tus estudios, trabajo o actividades cotidianas y debido a algún <u>PROBLEMA FÍSICO</u> (como estar cansada, no haber podido dormir bien, tener mareos…), durante las últimas 4 semanas…***

⑥ **¿Te ha <u>costado más mantener la atención</u> por algún problema físico?**

□ Nunca □ Casi nunca □ Algunas veces □ Bastantes veces □ Casi siempre □ Siempre

⑦ **¿Has tenido que <u>dejar de hacer algunas tareas</u> por algún problema físico?**

□ Nunca □ Casi nunca □ Algunas veces □ Bastantes veces □ Casi siempre □ Siempre



⑧ **¿Crees que, en las últimas 4 semanas, te <u>ha faltado confianza en tus propias capacidades</u>?**

□ Nunca □ Casi nunca □ Algunas veces □ Bastantes veces □ Casi siempre □ Siempre

⑨ **En las últimas 4 semanas, ¿has creído que tenías que <u>hacer las cosas perfectamente o no hacerlas</u>?**

☐ Nunca ☐ Casi nunca ☐ Algunas veces ☐ Bastantes veces ☐ Casi siempre ☐ Siempre

⑩ **En las últimas 4 semanas, ¿te has fijado <u>objetivos muy exigentes</u> y si no los has conseguido te has <u>sentido insatisfecha</u>?**

☐ Nunca ☐ Casi nunca ☐ Algunas veces ☐ Bastantes veces ☐ Casi siempre ☐ Siempre