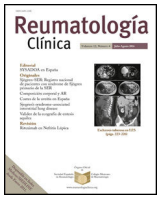
Reumatol Clin. 18 (2022) 236–245

Sociedad Española de Reumatología - Colegio Mexicano de Reumatología

Reumatología Clínica

www.reumatologiaclinica.org



Original

Validación de una versión en español del *Health Assessment Questionnaire-II* para la evaluación de la capacidad funcional en pacientes mexicanos con artritis reumatoide

Gabriel Horta-Baas*

Servicio de Reumatología, Hospital General Regional 1, Delegación Yucatán. Instituto Mexicano del Seguro Social, Mérida, Yucatán, México

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Recibido el 9 de agosto de 2020
Aceptado el 19 de noviembre de 2020
On-line el 10 de febrero de 2021

Palabras clave:
Capacidad funcional
Evaluación de la discapacidad
Clinimetría
Psicometría
Resultados reportados por el paciente
Artritis reumatoide

RESUMEN

Objetivo: Evaluar la validez de constructo, la fiabilidad y el funcionamiento de los ítems de una versión en español del *Health Assessment Questionnaire-II* (HAQ-II) para medir la capacidad funcional.
Métodos: Estudio transversal que incluyó a 496 pacientes con artritis reumatoide, distribuidos en 2 muestras. La validez de constructo se evaluó mediante el análisis factorial confirmatorio y la validez basada en la relación con otras variables. Para determinar la fiabilidad se empleó el coeficiente alfa de Cronbach (α) y omega de McDonald (ω). El funcionamiento de los ítems se analizó mediante el ajuste a diferentes modelos de la teoría de respuesta al ítem.
Resultados: El modelo de un factor presentó un mal ajuste en el análisis factorial confirmatorio; se realizó un análisis factorial exploratorio que señaló una estructura de 2 factores. El análisis factorial confirmatorio en la segunda muestra confirmó que el modelo de segundo orden presentó un buen ajuste a los datos. El factor general explicó más del 70% de la varianza. Los índices de fiabilidad mostraron una adecuada consistencia interna (α = 0,92-0,95; ω = 0,88-0,93). El 93% de las hipótesis contrastadas sobre la relación de las puntuaciones del HAQ-II con otras variables se confirmaron, lo cual demuestra su validez convergente, divergente y de grupos conocidos. El modelo de respuesta graduada multidimensional fue el que mejor predijo la interacción de las personas con los ítems.
Conclusión: La versión en español del HAQ-II presenta una adecuada validez y fiabilidad para la medición de la capacidad funcional en pacientes mexicanos con artritis reumatoide.

Validation of a Spanish Version of the Health Assessment Questionnaire-II to Assess Mexican Patients' Physical Function with Rheumatoid Arthritis

Keywords:
Functional status assessment
Disability evaluation
Clinimetrics
Psychometrics
Patient reported outcome measures
Rheumatoid arthritis

ABSTRACT

Objective: To evaluate the validity, reliability, and performance of the Health Assessment Questionnaire-II(HAQ-II) Spanish version questionnaire to measure physical function.
Methods: A cross-sectional study of 496 patients with rheumatoid arthritis, distributed in 2 samples. The construct validity was evaluated employing the confirmatory factor analysis and the validity based on the relationship with other variables. Cronbach's alpha (α) and McDonald's omega (ω) coefficient were used to determine reliability. Item performance was analysed by fitting different models of item response theory.
Results: The one-factor model presented a poor fit in the confirmatory factor analysis; an exploratory factor analysis was carried out, which suggested a 2-factor structure. The confirmatory factor analysis in the second sample confirmed that the second-order model had a good fit to the data. The general factor

* Autor para correspondencia.
Correo electrónico: gabho@hotmail.com

https://doi.org/10.1016/j.reuma.2020.11.005

explained more than 70% of the variance. The reliability indices showed adequate internal consistency (α = .92-.95; ω = .88-.93). Ninety-three percent of the contrasting hypotheses about the relationship of the HAQ-II scores with other variables were confirmed, demonstrating their convergent, divergent, and known group validity. The multidimensional graduated response model was the one that best predicted person's interaction with the items.
Conclusion: The Spanish version of the HAQ-II presents adequate validity and reliability for measuring Mexican patients' physical function with rheumatoid arthritis.

Introducción

La inflamación articular característica de la artritis reumatoide puede producir dolor y disminución de la fuerza muscular, lo cual conduce al deterioro de la capacidad funcional de la persona. La medición de la capacidad funcional puede realizarse por medio de un instrumento de medición (cuestionarios) o mediante la observación directa con pruebas de ejecución de una tarea específica asociada con la función (pruebas basadas en el rendimiento)[1] y constituye un desenlace relevante en el tratamiento y seguimiento de los pacientes con artritis reumatoide[2,3].

La forma habitual de valorar la capacidad funcional en las personas con artritis reumatoide es mediante cuestionarios autocumplimentados, teniendo en cuenta la percepción de los pacientes sobre sus habilidades físicas y la dificultad que la persona tiene para realizar sus actividades de la vida diaria. La percepción del paciente sobre su limitación funcional predice la discapacidad física, la necesidad de prótesis articular, la probabilidad de pérdida de empleo y de muerte prematura[2,4].

Los cuestionarios más empleados incluyen el *Health Assessment Questionnaire-Disability Index* (HAQ-DI), el *Health Assessment Questionnaire-II* (HAQ-II) y el *Multi-Dimensional Health Assessment Questionnaire* (MDHAQ)[5]. El cuestionario HAQ-DI es considerado el patrón de referencia para la evaluación de la limitación funcional[6,7]. Sin embargo, la extensión del HAQ-DI limita su uso rutinario; además, se ha demostrado que presenta algunos inconvenientes en sus propiedades psicométricas: presencia de efecto suelo, ausencia de linealidad y la posible mala interpretación de algunos ítems[4,8].

El HAQ-II se desarrolló para mejorar las propiedades psicométricas del HAQ-DI empleando métodos basados en la teoría de respuesta al ítem (TRI)[4]. Recientemente, el Colegio Americano de Reumatología recomendó al HAQ-II para la evaluación del funcionamiento físico en la práctica clínica diaria[5]. Existen estudios que demuestran la validez de las versiones en inglés[4,9], holandés[10] y español[11] del HAQ-II.

Es conocido que por la diversidad cultural existen diferencias entre el español de Europa, el español de América y entre las distintas variantes del español americano[12]. Estas diferencias pueden conducir a una interpretación errónea o a la incomprensión de un enunciado, sobre todo cuando presentan simultáneamente muchas diferencias fonéticas, morfosintácticas y léxicas[12]. Existen 2 versiones en español del HAQ-II desarrolladas para su uso en Argentina y Estados Unidos de América (EE. UU.)[11,13]. Waimann et al.[11] reportaron una adecuada fiabilidad y validez convergente de la versión en español de esta herramienta para su uso en Argentina. Después de la revisión bibliográfica no se encontró ningún estudio que reporte las propiedades psicométricas de la versión en español de EE. UU., pero se consideró que esta versión es la más idónea para su uso en la población mexicana por las siguientes razones: 1) los hablantes de origen mexicano representan casi 2/3 (63%) de los hispanohablantes estadounidenses[14]; 2) debido a que en los EE. UU. conviven personas hispanas de distintos orígenes, al momento de realizar la traducción de un instrumento se considera el empleo de un «español neutro» o «español internacional»; es decir, se elabora una versión lingüística que sea apropiada para más de un país[15,16]; 3) estudios previos demuestran la validez de la versión en español de EE. UU. del *Patient Activity Scale-II* (PAS-II) y del *Medical Outcome Study Pain Severity Scale* en la evaluación de pacientes mexicanos con artritis reumatoide[3,17,18] y 4) la traducción fue hecha por el grupo de autores que desarrollaron la versión original en inglés.

La adaptación transcultural de los instrumentos es útil cuando existe evidencia de adecuadas propiedades psicométricas en otras poblaciones y su propósito es asegurar que el instrumento presenta las mismas propiedades psicométricas del original para así garantizar la comparabilidad de los resultados[3]. Es decir, lo que se valida es la utilización y la interpretación de los resultados, no el instrumento *per se*. A la fecha no hay ningún estudio que evalúe las propiedades psicométricas de una versión en español del HAQ-II en población mexicana. Los objetivos del estudio son: 1) evaluar la validez de constructo; 2) evaluar la fiabilidad y 3) analizar el comportamiento y funcionamiento de los ítems del HAQ-II.

Material y métodos

Estudio transversal con un diseño para la validación de un cuestionario. La muestra del estudio incluyó pacientes con artritis reumatoide atendidos en la consulta externa de reumatología de 2 hospitales de segundo nivel de atención ubicados en el centro y sureste de México, seleccionados mediante un muestreo de casos consecutivos. Los pacientes están distribuidos en 2 muestras: 1) pacientes evaluados de marzo de 2018 a febrero de 2020 y 2) pacientes evaluados de marzo a septiembre de 2016; con la finalidad de efectuar en una de ellas el análisis factorial exploratorio (AFE) y en la otra el análisis factorial confirmatorio (AFC). Los criterios de inclusión fueron: cumplir los criterios de clasificación ACR/EULAR para AR[19] y tener una edad ≥ 16 años. Se excluyó a los pacientes que no sabían leer y a aquellos que no desearon completar los cuestionarios. El estudio cumplió con los estándares éticos de la Declaración de Helsinki y fue aprobado por el Comité de Ética del Instituto Mexicano del Seguro Social. Se obtuvo el consentimiento informado de todas las personas que participaron en el estudio.

El cálculo del tamaño de la muestra se basó en las recomendaciones de expertos en el tema. Para el análisis factorial, el tamaño de la muestra debe ser al menos 7 veces el número de ítems (es decir, 70 pacientes) con un mínimo de 100[20]. Con respecto al tamaño de la muestra para el análisis de la TRI, las guías actuales indican que una muestra de ≥500 pacientes se considera adecuada para los modelos logísticos de 2 parámetros[20].

Los participantes completaron una serie de cuestionarios en formato impreso: HAQ-II[13], HAQ-DI[15], *Rheumatoid Arthritis Quality of Life Scale* (RAQoL)[21], Escala Hospitalaria de Ansiedad y Depresión (EHAD)[22] y la escala de gravedad del dolor MOS (MOS-PSS)[17]. En caso de ser requerido, el mismo médico aclararó las dudas al respecto. Después de completar los cuestionarios, cada paciente fue evaluado clínicamente por el mismo reumatólogo y se calculó el grado de actividad con base en el *Clinical Disease Activity Index* (CDAI)[3] y con el *Disease Activity Score* 28 (DAS28)[3]. A diferencia

de los pacientes de la primera muestra, los pacientes de la segunda muestra carecían del resultado del RAQoL, HAQ-DI y del DAS28.

Instrumento

El HAQ-II consta de 10 ítems con una escala de respuesta de 4 categorías. La puntuación se obtiene mediante el promedio de los ítems y requiere por lo menos 8 ítems contestados. El rango de las puntuaciones va de 0 a 3, donde una puntuación más alta indica un mayor grado de limitación funcional. En este estudio se empleó la traducción al español de EE. UU., obtenido de la dirección electrónica del *National Data Bank of Rheumatic Diseases*[13]. Esta versión contiene un lenguaje simple y un vocabulario de uso común en nuestra población. En una prueba piloto previa a su uso en la muestra 2 (n = 30) se demostró su adecuada comprensión. La mayoría de los pacientes entendieron y respondieron sin dificultad. Sin embargo, se hicieron pequeños cambios con respecto a la original. En 2 preguntas se agregaron términos para clarificar el significado de las palabras: En el ítem «¿Sentarse y levantarse del inodoro?» se agregó la clarificación «(retrete, taza del baño)» y en el ítem «¿Abrir las puertas de un auto?» se agregó la clarificación «(coche)», que son términos empleados habitualmente en nuestra población. Por otra parte, en el ítem «¿Alcanzar y bajar un objeto de 2 kilos. Como 2 bolsas de azúcar desde una altura por encima de su cabeza?» se cambio la redacción a «¿Alcanzar y bajar un objeto de 2 kg, como una bolsa de azúcar, desde una altura por encima de su cabeza?», dado que en nuestra población la presentación de 2 kg es una presentación habitual para comprar el azúcar. El instrumento HAQ-II y la descripción de los otros instrumentos empleados en este estudio pueden encontrarse en el material suplementario.

Evaluación de la validez de constructo

La validez de constructo evalúa el grado en que el instrumento refleja la teoría del constructo que se desea medir (capacidad funcional); proporciona la evidencia de si la forma de interpretar las puntuaciones es correcta según la teoría y los constructos que se miden[23]. Para demostrar la validez de constructo del HAQ-II se analizó la validez basada en la estructura interna y la validez basada en la relación con otras variables.

Validez basada en la estructura interna

El análisis de la estructura interna evalúa si los ítems se ajustan a la dimensionalidad descrita al desarrollar el instrumento; es decir, analiza si la estructura interna del instrumento se mantiene invariante. El análisis factorial es una técnica estadística que permite evaluar la estructura interna de un instrumento, definir el número de factores y qué ítems se agrupan en cada factor. Con el análisis factorial se puede demostrar si la evidencia empírica permite aceptar el modelo basado en la teoría sobre el constructo que se evalúa o si se debe rechazar y plantear un nuevo modelo que explique el constructo medido. Existen 2 tipos: el AFE y el AFC. El AFE permite explorar, basado en los datos, los posibles constructos que explican las respuestas a los ítems de un instrumento. Una vez obtenido el número de factores, dependiendo del contenido de los ítems incluidos en cada factor, *a posteriori* se define el nombre del constructo obtenido. Por otro lado, el AFC precisa de una teoría que explique las respuestas a los ítems y que permita establecer las especificaciones del modelo; es decir, la construcción del modelo está basada en información *a priori*. El AFC comprueba si la estructura factorial es consistente (presenta un buen ajuste a los datos) o no con la estructura teórica y se emplea para demostrar la validez de constructo de los modelos obtenidos en el AFE[24].

El procedimiento de la evaluación de la estructura interna del HAQ-II requirió de 3 etapas: 1) se utilizó en la primera muestra (n = 343) el AFC, evaluando el modelo de un factor descrito en la versión original en inglés[4]; 2) dado que la versión en español presentó un mal ajuste al modelo de un factor, se analizó la posibilidad de un nuevo modelo con base en el AFE y 3) la segunda muestra (n = 153) se utilizó para validar el modelo obtenido en el AFE. Dos modelos fueron evaluados con el AFC: 1) modelo de 2 factores correlacionados; 2) modelo de 2 factores de primer orden y un factor de segundo orden.

Validez basada en la relación con otras variables

La relación con otras variables que miden el mismo constructo (validez convergente) o diferentes constructos (validez divergente) son relevantes cuando las puntuaciones del instrumento se usan para estimar el nivel de los pacientes en un constructo, dado que proporcionan información de la magnitud en que estas relaciones son acordes con el constructo en el que se basa la interpretación de los resultados. Se espera que la correlación entre los instrumentos que miden el mismo constructo sea mayor que entre los instrumentos que miden constructos diferentes. La validez de grupos conocidos se demuestra cuando las puntuaciones del instrumento discriminan entre los grupos de pacientes que, teóricamente, se espera que sean diferentes en el constructo medido.

Para la evaluación de la validez convergente se contrastaron las siguientes hipótesis:

1. La correlación entre el HAQ-II y el HAQ-DI[15] debe ser alta ($\rho \geq 0{,}7$), dado que evalúan el mismo constructo.
2. Debido a que son constructos relacionados pero conceptualmente distintos, la puntuación del HAQ-II deberá tener una correlación moderada o mayor ($\rho \geq 0{,}5$) con la gravedad del dolor (medida con el cuestionario MOS-PSS[17]), la fatiga (medida con una escala visual analógica de 0 a 10), la evaluación global del paciente (medida con una escala visual analógica de 0 a 10), la actividad de la enfermedad (medida con los índices CDAI y DAS28[3]) y la calidad de vida (medida con el cuestionario RAQoL[21]).

Para la evaluación de la validez divergente se contrastó la siguiente hipótesis:

1. La puntuación del HAQ-II presentará una baja correlación ($0{,}3 > \rho < 0{,}49$)[4] con la puntuación de las subescalas de ansiedad y depresión de la EHAD, dado que son constructos distintos.

Para la evaluación de la validez de grupos conocidos se contrastó la siguiente hipótesis:

1. Los pacientes con actividad moderada a severa con base en la puntuación del CDAI (CDAI > 10) presentarán una puntuación del HAQ-II significativamente mayor que los pacientes en remisión clínica y baja actividad de la enfermedad (CDAI $\leq$ 10).

Para demostrar la evidencia de la validez de constructo se consideró necesario confirmar el 75% de las hipótesis planteadas *a priori*[25].

Evaluación de la fiabilidad

La fiabilidad evalúa la precisión con la cual un instrumento mide un constructo; es decir, proporciona mediciones libres de error[23]. La consistencia interna mide la homogeneidad de los ítems de un instrumento indicando la relación entre ellos y es el método más empleado para medir la fiabilidad de un instrumento. Dentro del análisis de la TRI, el coeficiente de fiabilidad marginal proporciona una estimación de la precisión con que mide un instrumento. El

coeficiente de fiabilidad marginal se interpreta como la proporción de varianza en la puntuación observada que se debe a la puntuación verdadera[26].

Análisis basado en la teoría de respuesta al ítem

La TRI agrupa una serie de modelos que, por medio de funciones matemáticas, describen la probabilidad que tiene un paciente de seleccionar una determinada respuesta a un ítem de acuerdo con su nivel del rasgo latente (capacidad funcional). La TRI permite mediciones invariantes más allá de los ítems que componen el instrumento y provee información de la exactitud con que se mide el constructo en función del nivel del rasgo latente (θ). En la TRI el ítem se considera la unidad de análisis y el nivel del rasgo latente que presenta el paciente se estima a partir del patrón de respuesta obtenido del conjunto de ítems[27]. El número de rasgos latentes que intervienen al contestar un ítem establece si el modelo es unidimensional (un rasgo latente) o multidimensional (más de un rasgo latente). Los modelos de la TRI vinculan el nivel θ del paciente con los parámetros del ítem (discriminación y dificultad) y la probabilidad de seleccionar una respuesta. El parámetro de discriminación (α) mide la fuerza de la relación entre el ítem y el rasgo latente que se mide[28]. Los parámetros de dificultad (parámetros β) se interpretan como desviaciones estándar que muestran el rango del rasgo latente cubierto por el ítem. Cuanto mayor sea el parámetro β, mayor será el nivel de limitación funcional que una persona debe tener para seleccionar esa opción de respuesta[28]. Las pruebas de bondad de ajuste permiten evaluar hasta qué punto el modelo representa los datos observados[27]. Si el modelo se ajusta a los datos empíricos, se puede suponer que el modelo representa de forma apropiada la relación entre el rasgo latente y la probabilidad de que el paciente seleccione determinada respuesta al ítem.

Análisis estadístico

El análisis descriptivo de las variables categóricas (características de los pacientes y de los ítems) y el efecto suelo-techo (pacientes que obtenían la puntuación mínima y máxima posible, respectivamente) se presentan como el número de casos (n) y su porcentaje (%). En el caso de las variables continuas se presentan como media y desviación estándar (media ± DE) o mediana (rango intercuartil), según corresponda. Para la comparación entre las puntuaciones del factor 1 y del factor 2 se empleó la prueba t para muestras relacionadas. La concordancia entre el HAQ-DI y el HAQ-II se evaluó con el coeficiente de correlación y concordancia de Lin (CCC) y por el método de Bland-Altman[29].

Análisis factorial. El AFC se realizó utilizando correlaciones policóricas y el procedimiento de estimación de los parámetros empleado fue el de mínimos cuadrados ponderados diagonalmente (DWLS)[30]. Los criterios que se emplearon para determinar un buen ajuste del modelo son: una razón de ji cuadrado (χ^2) sobre los grados de libertad (CMIN/DF) < 3, el error cuadrático medio de aproximación (RMSEA) < 0,06, el residuo cuadrático medio estandarizado (SRMR) ≤ 0,08, el índice de ajuste comparativo (CFI) ≥ 0,95 y el índice de Tucker Lewis (TLI) ≥ 0,95. Un valor RMSEA ≤ 0,08 se considera un ajuste aceptable y un valor > 0,10 se considera un mal ajuste[18]. En el AFE, el número apropiado de factores que extraer se determinó con el análisis paralelo, los parámetros se estimaron con el método de residuales mínimos (minres) y se empleó una rotación oblimin.

Validez basada en la relación con otras variables. Se calculó el coeficiente de correlación de Spearman (ρ) entre las puntuaciones del HAQ-II y las puntuaciones de los otros instrumentos. Se empleó la prueba de Kruskal-Wallis y la prueba de Dunn para saber entre qué grupos las puntuaciones del HAQ-II diferían significativamente según el grado de actividad medido por el CDAI[31].

Fiabilidad. La consistencia interna se estimó con los coeficientes α de Cronbach y ω de McDonald. Estos coeficientes tienen la misma interpretación: un valor entre 0,70 y 0,95 se considera apropiado[32].

Análisis basado en la TRI. Los modelos de la TRI evaluados fueron: el modelo de crédito parcial (MCP), el modelo de crédito parcial generalizado (MCPG) unidimensional y multidimensional (MCPGM), modelo de respuesta graduada unidimensional (MRG) y multidimensional (MRGM). El ajuste del modelo a los datos a nivel del ítem se evaluó con el estadístico S-X^2, un valor de *p* ajustado mediante el método de Benjamini-Hochberg < 0,05 indica un mal ajuste a nivel del ítem[18]. Se consideró un buen ajuste global del modelo si el valor de *p* del estadístico de información limitada (M2*) era > 0,05, RMSEA < 0,089 y SRMR < 0,05[18]. Los índices G2-LD y Q3 se emplearon para evaluar el supuesto de independencia local de los ítems[30]. Se compararon los resultados de los modelos a partir de los índices de bondad de ajuste, la razón de verosimilitudes (*likelihood ratio test*) y la prueba de Vuong[33]. Para evaluar el ajuste de las personas (*person fit*) se empleó el estadístico Zh[34]. El estadístico Zh se emplea para identificar a las personas cuya respuesta a los ítems no es consistente con el nivel de su rasgo latente. Un valor del Zh mayor de ± 2 refleja a las personas con un patrón de respuesta «atípico» o «inconsistente»[34].

El análisis estadístico se realizó en el programa R (R Core Team, 2017, Viena, Austria). Los paquetes empleados en el análisis de los datos fueron: *psych* para el análisis descriptivo y el AFE; *lavaan* y *semplot* para el AFC; *psych* y *semtools* para el análisis de fiabilidad y *mirt* para la calibración de los modelos de la TRI.

Resultados

Características generales de los participantes

Cumplieron los criterios de inclusión 508 pacientes, de los cuales 496 (97,63%) completaron todos los ítems del HAQ-II. Se eliminó a los participantes con ítems incompletos; la mayoría dejó un ítem sin respuesta; solo un paciente no respondió 3 ítems. Los ítems que presentaron el mayor número de valores perdidos fueron el ítem 6 (n = 2) y el 10 (n = 2). La muestra total (n = 496) se conformó principalmente por mujeres (87,5%), con una edad media de 49,8 años (rango: 16-79) y una escolaridad promedio de 9 años (tabla 1). El HAQ-DI presentó una mayor frecuencia del efecto suelo (11,1%; 38 de 343) que el HAQ-II (8,2%; 28/343). Ni el HAQ-II ni el HAQ-DI presentaron el efecto techo.

Validez basada en la estructura interna

Los resultados de la calibración de los ítems en la primera muestra se presentan en la tabla 2. Los resultados del AFC indicaron una pobre adecuación al modelo de un factor (RMSEA > 0,10), por lo cual se realizó un AFE a los datos, que demostró que una solución de 2 factores correlacionados explica las respuestas a los ítems del HAQ-II. El AFC en la segunda muestra demostró que el modelo de 2 factores correlacionados y el modelo de segundo orden presentaron un ajuste aceptable a los datos (RMSEA = 0,08; SRMR = 0,06). La reespecificación del modelo, incluyendo la correlación entre los residuales de los ítems 8 y 9, ocasionó un buen ajuste del modelo a los datos (fig. 1).

Validez basada en la relación con otras variables

El 93% de las hipótesis contrastadas sobre la relación de las puntuaciones del HAQ-II con otras variables se confirmaron, lo cual demuestra su validez convergente, divergente y de grupos conocidos. La puntuación total del HAQ-II se calculó como el promedio del factor 1 y del factor 2. Los coeficientes de correlación de las

Tabla 1
Características generales de los participantes en el estudio

	Muestra 1 (n = 343)	Muestra 2 (n = 153)	Total (n = 496)
Edad	49,28 ± 11,50	51,01 ± 11,63	49,81 ± 11,56
Escolaridad	9,77 ± 4,13	7,9 ± 3,57	9,22 ± 4,06
Health Assessment Questionnaire-II (modelo de un factor)	0,8 (1)	0,8 (0,9)	0,8 (0,9)
Health Assessment Questionnaire-II (modelo de 2 factores)			
Factor 1	0,6 (1)	0,6 (0,8)	0,6 (1)
Factor 2	0,8 (1,2)	0,8 (1)	0,8 (1,2)
Total	0,8 (1)	0,8 (0,9)	0,8 (0,9)
Health Assessment Questionnaire-II (2 factores, valor theta, θ)			
Factor 1	0,01 ± 0,95	−0,10 ± 0,96	−0,01 ± 0,94
Factor 2	0,02 ± 0,93	−0,08 ± -0,10	−0,01 ± 0,95
Health Assessment Questionnaire-Disability Index (HAQ-DI)	1 (1,12)	-	-
Clinical Disease Activity Index (CDAI)	8 (15,5)	6,25 (12,5)	7 (14,5)
Disease Activity Score 28 (DAS28)	3,61 ± 1,45	-	-
Rheumatoid Arthritis Quality of Life Scale (RAQoL)	10 (14)	-	-
Escala de gravedad del dolor MOS (MOS-PSS)	51,42 (40)	54,28 (28,57)	54,28 (38,57)
Escala visual analógica de fatiga	3,5 (5,5)	3,2 (5,5)	3,5 (5,5)
Escala visual analógica de la evaluación global del paciente	3,5 (5,5)	3 (5)	3 (5,5)
Escala hospitalaria de ansiedad y depresión			
Ansiedad	7 (5)	9 (7)	7 (5)
Depresión	6 (5)	8 (5)	6 (5)

Tabla 2
Resultados del análisis factorial confirmatorio y del análisis factorial exploratorio en los pacientes de la muestra 1

Health Assessment Questionnaire-II (HAQ-II)		Clasificación Internacional del Funcionamiento (CIF)			AFC Modelo de 1 factor		AFE Modelo de 2 factores	
					Carga factorial (error estándar)	R^2	Carga factorial	
Ítem	Descripción	Componente de la CIF	Código CIF	Categoría CIF			Factor 1	Factor 2
1	Levantarse de una silla sin ayudarse con los brazos	Movilidad	D4103	Cambiar las posturas corporales básicas/sentarse	0,82 (0,00)	0,68	0,98	0,12
2	Caminar fuera de su casa, sobre un terreno plano	Movilidad	D450	Andar	0,80 (0,04)	0,64	0,66	0,16
3	Sentarse y levantarse del inodoro (retrete, taza del baño)	Movilidad	D4103	Cambiar las posturas corporales básicas/sentarse	0,86 (0,03)	0,74	0,86	0,04
4	Alcanzar y bajar un objeto de 2 kilos, como una bolsa de azúcar, desde una altura por encima de su cabeza	Movilidad	D4452	Uso de la mano y el brazo/alcanzar	0,82 (0,03)	0,68	0,32	0,54
5	Abrir las puertas de un auto (coche)	Movilidad	D4450	Uso de la mano y el brazo/tirar-halar	0,81 (0,03)	0,66	0,38	0,43
6	Limpiar un patio	Vida doméstica	D640	Realizar los quehaceres de la casa	0,85 (0,03)	0,73	0,44	0,47
7	Esperar de pie en una fila durante 15 minutos	Movilidad	D4154	Mantener la posición del cuerpo/permanecer de pie	0,84 (0,03)	0,71	0,47	0,42
8	Mover objetos pesados	Movilidad	D449	Llevar, mover y usar objetos	0,97 (0,03)	0,94	0,01	0,99
9	Sostener objetos pesados	Movilidad	D440	Uso fino de la mano/agarrar	0,96 (0,03)	0,92	0,02	0,99
10	Subir rápidamente las escaleras	Movilidad	D455	Desplazarse por el entorno	0,81 (0,03)	0,66	0,58	0,24
Varianza extraída media (AVE). La varianza del constructo que se pueda explicar a través de los ítems					0,74	-		-
Estadísticos de bondad de ajuste de los modelos del AFC	Razón de χ^2 sobre los grados de libertad (CMIN/ DF)				5,85			-
	Error cuadrático medio de aproximación (RMSEA) (Intervalo de confianza 90%)				0,11 (0,10-0,13)		0,05 (0,01-0,09)	
	Índice de ajuste comparativo (CFI)				0,99			-
	Índice de Tucker (TLI)				0,99		0,99	
	Residuo cuadrático medio estandarizado (SRMR)				0,06		0,02	
Coeficientes de fiabilidad	α de Cronbach				0,96		0,93	0,95
	ω de MacDonald				0,94		0,89	0,93

AFC: análisis factorial confirmatorio; AFE: análisis factorial exploratorio.

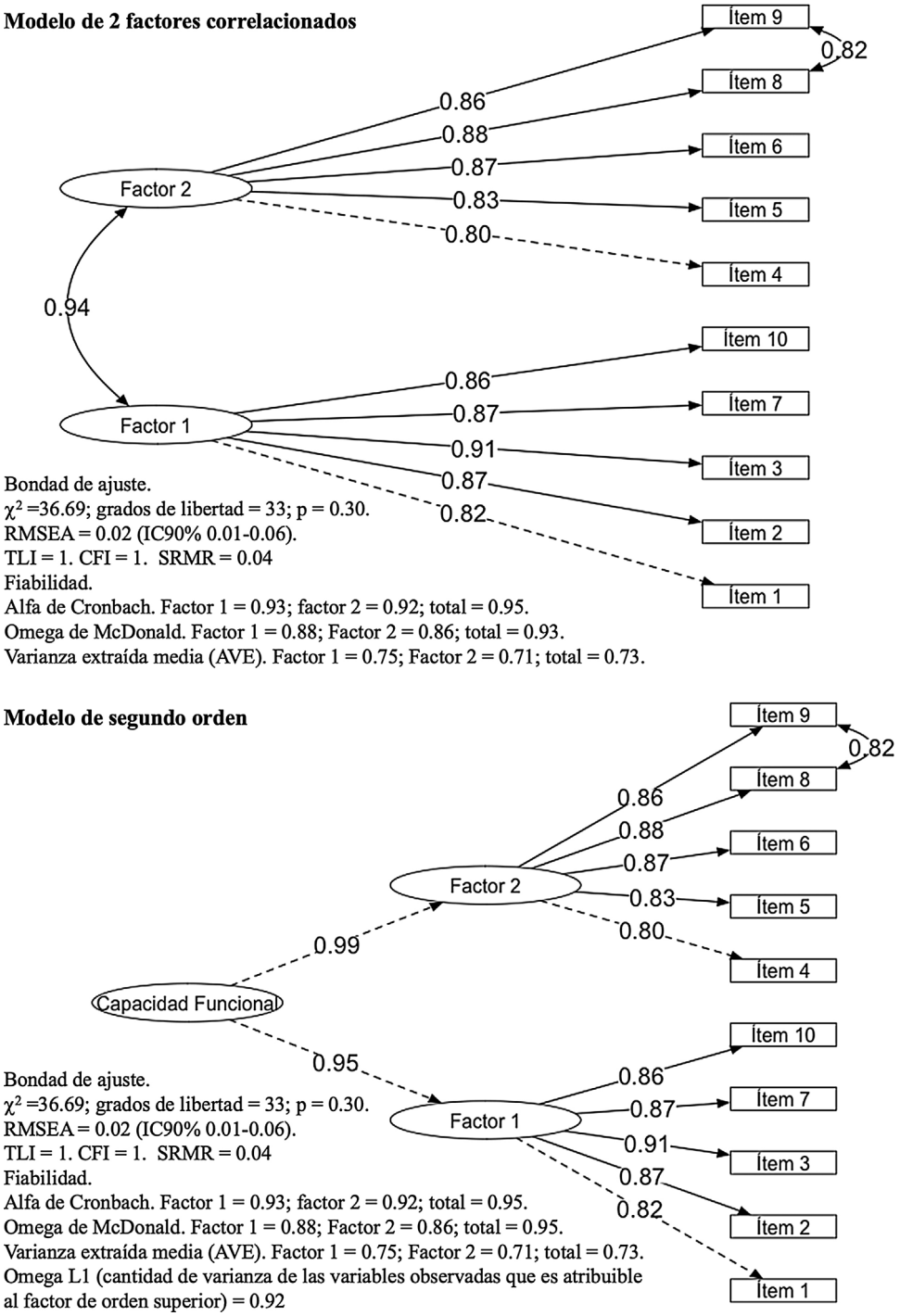


Figura 1. Representación (*path diagram*) de los 2 modelos que mejor describen la estructura interna de la versión en español del HAQ-II.

puntuaciones del HAQ-II con otras variables se presentan en la tabla 3.

La puntuación del HAQ-II fue capaz de diferenciar a los pacientes con actividad inflamatoria de los pacientes con remisión/baja actividad de la enfermedad. La puntuación del HAQ-II fue significativamente mayor de acuerdo con el grado de actividad de la enfermedad por el CDAI ($p < 0{,}001$). Con excepción de la comparación de los grupos de actividad moderada frente a los de actividad grave, todas las comparaciones fueron significativas ($p < 0{,}005$) en el análisis *post hoc*.

Fiabilidad

El resultado de los coeficientes α de Cronbach (rango de 0,92-0,95) y ω de McDonald (rango de 0,88-0,93) demostraron una adecuada consistencia interna (tabla 2 y fig. 1). El coeficiente de fiabilidad marginal del factor 1 y del factor 2 fue de 0,89 y 0,92,

Tabla 3
Correlación entre las puntuaciones del *Health Assessment Questionnaire-II* y *Health Assessment Questionnaire-Disability Index* con otras variables clínicas

	HAQ-DI	Actividad (CDAI)	Actividad (DAS28)	Calidad de vida (RAQoL)	Dolor (MOS-PSS)	Fatiga (EVA)	Evaluación del paciente (EVA)	Ansiedad (EHAD)	Depresión (EHAD)
HAQ-II (un factor)	0,87 (0,84-0,89)	0,56 (0,49-0,62)	0,51 (0,41-0,60)	0,72 (0,66-0,77)	0,64 (0,57-0,71)	0,68 (0,62-0,73)	0,65 (0,59-0,71)	0,33 (0,22-0,43)	0,40 (0,29-0,49)
HAQ-II (2 factores, puntuación promedio de los ítems)									
Factor 1	0,81 (0,77-0,84)	0,52 (0,45-0,58)	0,48 (0,37-0,57)	0,65 (0,58-0,71)	0,61 (0,54-0,67)	0,68 (0,58-0,70)	0,62 (0,55-0,68)	0,30 (0,19-0,40)	0,35 (0,25-0,45)
Factor 2	0,84 (0,80-0,87)	0,54 (0,47-0,60)	0,48 (0,37-0,57)	0,70 (0,64-0,75)	0,59 (0,52-0,66)	0,63 (0,56-0,69)	0,62 (0,55-0,68)	0,32 (0,21-0,42)	0,40 (0,29-0,49)
Resumen (Promedio factor 1 y 2)	0,87 (0,85-0,89)	0,56 (0,49-0,62)	0,51 (0,41-0,60)	0,72 (0,66-0,77)	0,64 (0,57-0,69)	0,68 (0,62-0,73)	0,65 (0,59-0,71)	0,33 (0,22-0,43)	0,40 (0,30-0,49)
HAQ-II (2 factores, valor theta θ)									
Factor 1	0,85 (0,82-0,87)	0,55 (0,48-0,61)	0,47 (0,36-0,56)	0,70 (0,63-0,75)	0,64 (0,57-0,69)	0,64 (0,57-0,70)	0,62 (0,55-0,68)	0,32 (0,21-0,42)	0,38 (0,28-0,47)
Factor 2	0,84 (0,80-0,86)	0,52 (0,44-0,60)	0,51 (0,41-0,60)	0,70 (0,64-0,75)	0,60 (0,52-0,66)	0,67 (0,61-0,73)	0,65 (0,58-0,71)	0,31 (0,20-0,41)	0,40 (0,30-0,50)
HAQ-DI	1	0,55 (0,47-0,62)	0,52 (0,42-0,61)	0,73 (0,67-0,78)	0,66 (0,59-0,72)	0,64 (0,58-0,70)	0,61 (0,54-0,67)	0,34 (0,22-0,44)	0,41 (0,30-0,50)

n = 343. Correlación de Spearman (intervalo de confianza 95%),
CDAI: *Clinical Disease Activity Index*; DAS28: *Disease Activity Score 28*; EHAD: escala hospitalaria de ansiedad y depresión; EVA: escala visual analógica; MOS-PSS: escala de gravedad del dolor MOS; RAQoL: *Rheumatoid Arthritis Quality of Life Scale*.

respectivamente, lo cual demuestra que se obtienen mediciones precisas según este indicador de la TRI.

Modelos de la teoría de respuesta al ítem

Para obtener la mayor precisión de los parámetros estimados, el análisis basado en la TRI se realizó con la muestra total de pacientes (n = 496). Los resultados del ajuste para los modelos evaluados se presentan en la tabla 4. El modelo de respuesta graduada multidimensional (MRGM) de 2 factores correlacionados fue el que presentó el mejor ajuste a los datos y fue significativamente mejor que el MCPGM ($p = 0,03$). Un par de ítems (ítems 8 y 9) mostraron un mal ajuste a nivel del ítem con base en los estándares establecidos. Sin embargo, la eliminación de cualquiera de estos ítems produjo un mal ajuste global del modelo. Ninguno de los ítems presentó dependencia local.

Parámetros de los ítems. Las estimaciones de los parámetros y los estadísticos de ajuste de cada uno de los ítems se presentan en la tabla 5. El rango de los parámetros de discriminación fue de 2,21 a 7,36, lo cual indica que los ítems permiten discriminar bien entre las personas con alta y baja capacidad funcional. El ítem con más capacidad de distinguir entre pacientes fue el ítem 8 «Mover objetos pesados». Por el contrario, el ítem 5 «abrir las puertas de un auto» fue el que demostró la menor discriminación. Todos los parámetros de dificultad (β) de los ítems presentaron un ordenamiento creciente, lo que indica que es necesario tener una puntuación más alta en la limitación funcional a medida que se aumenta en la categoría de respuesta seleccionada. Los ítems «caminar fuera de su casa» y «abrir las puertas de un auto» fueron los ítems «más difíciles», es decir, que requieren un alto nivel de limitación funcional para ser seleccionados. Por el contrario, los ítems «mover objetos pesados» y «sostener objetos pesados» fueron los ítems «más fáciles», es decir, se requiere que los pacientes presenten una limitación funcional leve para que sean seleccionados, lo que indica que en estos ítems es donde las personas mostraron más frecuentemente su limitación funcional.

Parámetros de las personas. Las estimaciones de la capacidad funcional (θ) de los pacientes en el primer factor se distribuyeron en valores *logits* de −1,67 y 2,73; y entre −1,76 y 2,58 en el segundo factor. El análisis del ajuste de las personas mostró que el 95,37% de los patrones de respuesta se ajustaron al modelo (valor del estadístico Zh entre −2 y 2), y únicamente el 1,61% de las personas presentaron un patrón de respuestas «atípicas» o «inconsistentes», con un valor del estadístico Zh < −3.

Interpretación de los resultados

La correlación entre las puntuaciones obtenidas mediante el promedio de los ítems y el valor θ obtenido en el MRGM fue de 0,95 (IC 95%: 0,95-0,96) para el factor 1 y de 0,96 (IC 95%: 0,95-0,97) para el factor 2. La puntuación promedio de los ítems del factor 2 fue significativamente mayor que la del factor 1 en la muestra 1 (4,09 frente a 3,03; $p < 0,001$) y en la muestra 2 (3,78 frente a 2,74; $p < 0,001$). La magnitud y dirección de la correlación entre el HAQII y el HAQ-DI con las variables clínicas estudiadas fueron similares (tabla 5). La concordancia entre las puntuaciones del HAQ-II y del HAQ-DI presentó un CCC = 0,87 (IC 95%: 0,84-0,89) y el sesgo estimado por el método de Bland-Altman fue de 0,04 (límites de concordancia: −0,62, 0,72).

Discusión

Los resultados del estudio demuestran que la estructura interna de la versión en español del HAQ-II está constituida por 2 factores. Los modelos de 2 factores correlacionados y el modelo de segundo orden presentaron niveles equivalentes de ajuste. Sin embargo, la alta correlación entre los 2 factores, obtenida en el AFC, indica la presencia de un factor general. El modelo de segundo orden proporciona una explicación parsimoniosa de las correlaciones entre los factores de primer orden[35]. Los resultados del estudio muestran que el HAQ-II está constituido por un factor general que mide la capacidad funcional indirectamente (a través de los ítems) y 2 factores relacionados pero distintos, que miden características específicas de la capacidad funcional.

El análisis de la estructura factorial incide en la interpretación de la puntuación de un instrumento y puede revelar que la puntuación debe separarse en varias puntuaciones[35]. A partir del modelo de segundo orden se pueden obtener 3 puntuaciones: una por cada factor y una total. Debido a que el factor general (capacidad funcional) explica la mayor parte de la varianza observada en la respuesta a los ítems (>0,7), se puede considerar que el HAQ-II representa esencialmente un instrumento unidimensional[36]. En la práctica

Tabla 4
Resultados del ajuste de los modelos de la teoría de respuesta al ítem (TRI) al HAQ-II y comparación de su bondad de ajuste

Modelos	M2*(*p* valor)	RMSEA (IC 95%)	SRMR	TLI	AIC
Unidimensionales					
Modelo de crédito parcial (MCP; Rasch)	<0,001	0,12 (0,10-0,13)	0,06	0,95	8.261,62
Modelo de crédito parcial generalizado (MCPG)	<0,001	0,09 (0,07-0,11)	0,05	0,97	8.229,71
Modelo de respuesta graduada (MRG)	<0,001	0,10 (0,08-0,12)	0,05	0,96	8.209,71
Multidimensionales					
Modelo de crédito parcial generalizado multidimensional (MCPGM)	0,005	0,04 (0,02-0,07)	0,05	0,99	8.093,53
Modelo de respuesta graduada multidimensional (MRGM)	0,14	0,02 (0,01-0,05)	0,05	0,99	8.069,96

Comparación de modelos:
MCPG en comparación con el MCPGM: prueba de razón de verosimilitudes (LR Test) $p < 0,001$, prueba de Vough $p < 0,001$.
MRG en comparación con el MRGM: LR Test $p < 0,001$, prueba de Vough $p < 0,001$.
MCPGM en comparación con el MRGM: LR Test $p < 0,001$, prueba de Vough $p = 0,03$.
AIC: criterio de información de Akaike; IC 95%: intervalo de confianza del 95%; RMSEA: error medio cuadrático de aproximación; SRMR: raíz media cuadrática residual; TLI: Índice de Tuker-Lewis.

Tabla 5
Parámetros estimados con el modelo de respuesta graduada multidimensional (MRGM), ajuste de los ítems y coeficiente de fiabilidad marginal de la TRI

Ítem	Descripción	Discriminación A (MDISC)	Dificultad (MDIFF)			Ajuste a nivel del Ítem	Coeficiente de fiabilidad marginal
			B1	B2	B3	S-χ^2 (*p*)	
Dimensión (factor) 1							
1	Levantarse de una silla sin ayudarse con los brazos	2,74	0,008	1,411	2,264	0,38	0,89
2	Caminar fuera de su casa, sobre un terreno plano	2,83	0,299	1,557	2,649	0,38	
3	Sentarse y levantarse del inodoro (retrete, taza del baño)	3,56	0,107	1,329	2,478	0,11	
7	Esperar de pie en una fila durante 15 minutos	3,01	−0,503	0,745	1,667	0,38	
10	Subir rápidamente las escaleras	2,92	−0,612	0,596	1,392	0,89	
Dimensión (factor) 2							
4	Alcanzar y bajar un objeto de 2 kilos, como una bolsa de azúcar desde una altura por encima de su cabeza	2,55	−0,400	0,927	1,771	0,54	0,92
5	Abrir las puertas de un auto (coche)	2,21	0,267	1,624	2,568	0,38	
6	Limpiar un patio	2,91	−0,336	1,003	1,955	0,48	
8	Mover objetos pesados	7,36	−1,080	0,274	0,906	0,005	
9	Sostener objetos pesados	6,37	−1,025	0,257	0,887	0,005	

Datos obtenidos del análisis de la muestra total (n = 496).
El parámetro A representa el parámetro de discriminación multidimensional (*MDISC*).
Los parámetros B representan la dificultad multidimensional (MDIFF). Los valores de B positivos altos indican ítems difíciles (es decir, aquellos que requieren valores altos de limitación funcional para producir una probabilidad de una respuesta correcta mayor que 0,5). Los valores B bajos y más negativos indican ítems con una alta probabilidad de respuesta correcta para las personas con bajos niveles de limitación funcional.
Ajuste global del modelo a los datos: M2*; $p = 0,14$; RMSEA 0,02 (IC 95%: 0-0,05); SRMR = 0,05; TLI = 0,99.

clínica, el empleo de la puntuación total podría ser la puntuación más útil.

Con base en el análisis del contenido de los ítems (tabla 2) incluidos en cada uno de los factores, a partir de la clasificación internacional del funcionamiento (CIF)[37], se consideró que el factor 1 mide «actividades con predominio del funcionamiento de las extremidades inferiores» y el factor 2 mide las «actividades con predominio del funcionamiento de las extremidades superiores». Debido a que la artritis reumatoide afecta de modo predominante a las articulaciones pequeñas y medianas de las extremidades superiores, existe una justificación teórica para considerar que la limitación funcional en los pacientes con esta enfermedad sea mayor en las extremidades superiores que en las inferiores. Para el buen ajuste a los datos fue necesario especificar una correlación entre los errores de los ítems 8 y 9, lo cual indica que ambos ítems presentan una varianza única compartida. Este hallazgo fue descrito previamente[10,11] y tiene sustento teórico. La correlación entre los ítems «mover objetos pesados» y «levantar objetos pesados» puede ser explicada por ser 2 tareas muy similares, es decir, las personas generalmente intentan levantar objetos para moverlos.

La magnitud y dirección de las correlaciones entre la puntuación total de la versión en español del HAQ-II con otras variables es consistente con las correlaciones publicadas en estudios previos: con la intensidad del dolor de 0,46-0,61[4,10,11], 0,55 con la fatiga[4], 0,55 con el DAS28[11], 0,44 con los síntomas depresivos[4], 0,38 con los síntomas de ansiedad[4] y de 0,91-0,92 con el HAQ-DI[4,10,11]. Estos resultados demuestran su validez basada en la relación con otras variables.

El análisis basado en la TRI tiene el objetivo de producir puntuaciones fiables que se puedan usar para hacer inferencias válidas sobre los examinados[38]. Los diferentes tipos de modelos representan diferentes perspectivas teóricas[39]. Un modelo es útil en la medida en que proporciona predicciones razonables de las probabilidades de selección de una respuesta al ítem y la ubicación de la persona en la escala, simultáneamente[39]. Dos estudios previos reportaron un buen ajuste de los datos al modelo

Rasch, lo cual implica que todos los ítems presentan la misma discriminación y miden una sola dimensión[4,10]. En otro, estudio se demostró un adecuado ajuste al MCPG[9], que permite variar la discriminación entre los ítems.

En este estudio examinamos el ajuste a diferentes tipos de modelos de la TRI. A diferencia de los estudios previos, los resultados muestran que ni los modelos unidimensionales (MCPG y MRG) ni el modelo Rasch (MCP) representaron con precisión la interacción ítems-personas en nuestra muestra. El MRGM de 2 dimensiones representó con mayor precisión la interacción entre los pacientes y los ítems del HAQ-II. Es decir, los ítems son sensibles a las diferencias en las actividades que requieren predominantemente el funcionamiento de las extremidades superiores de las actividades que requieren predominantemente el funcionamiento de las extremidades inferiores y que las diferencias en la limitación funcional de las personas requirieron de 2 dominios para representarse con precisión. El coeficiente de fiabilidad marginal de las estimaciones de la puntuación de las dimensiones[36] demuestra que el HAQ-II ordena de forma apropiada a los encuestados a lo largo del nivel del rasgo latente.

La alta correlación entre las puntuaciones estimadas por el promedio de los ítems y el valor θ obtenido del MRGM indica que el empleo del promedio de los ítems proporciona puntuaciones similares del rasgo latente. La concordancia entre las puntuaciones del HAQ-II y el HAQ-DI en nuestro estudio coinciden con lo reportado por Wolfe et al. (CCC = 0,902; sesgo = 0,02)[4]. Esto señala que ambas escalas evalúan constructos similares, pero no son intercambiables a nivel individual[4,10].

Las limitaciones que se deben considerar al interpretar el estudio son: debido al diseño del estudio algunas propiedades psicométricas no fueron evaluadas (fiabilidad test-retest, validez predictiva y sensibilidad al cambio); el estudio no incluye un porcentaje significativo de pacientes con baja alfabetización, en los cuales se puede presentar una mayor dificultad en la comprensión de los ítems; la muestra no proviene de la población general y no se evaluó si los pacientes tienen una preferencia por el HAQ-II en comparación con el HAQ-DI. Sin embargo, la escolaridad de los participantes es equivalente al promedio estimado para la población mexicana y se empleó una muestra de pacientes que constituyen la población diana para el uso del HAQ-II.

Conclusión

La versión en español del HAQ-II presenta un grado adecuado de validez y fiabilidad para la medición de la capacidad funcional en pacientes mexicanos con artritis reumatoide.

Financiación

La presente investigación no ha recibido ayudas específicas provenientes de agencias del sector público, sector comercial o entidades sin ánimo de lucro.

Conflicto de intereses

El autor declara no tener ningún conflicto de intereses.

Anexo. Material adicional

Se puede consultar material adicional a este artículo en su versión electrónica disponible en doi:10.1016/j.reuma.2020.11.005.

Bibliografía

1. Horta-Baas G, Vargas-Mena R, Alejandre E, Pelaez-Ballestas I, Romero-Figueroa MD, Queipo G. Psychometric properties of the 12-item Knee injury and Osteoarthritis Outcome Score (KOOS-12) Spanish version for people with knee osteoarthritis. Clin Rheumatol. 2020.
2. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum. 1993;36:729–40.
3. Horta-Baas G, Perez Bolde-Hernandez A, Hernandez-Cabrera MF, Vergara-Sanchez I, Romero-Figueroa MDS. Evaluación de la actividad de la artritis reumatoide en la atención clínica habitual. Concordancia entre la autoclinimetría y la evaluacion clínica con los índices de actividad: DAS28, CDAI y SDAI. Med Clin (Barc). 2017;149:293–9.
4. Wolfe F, Michaud K, Pincus T. Development and validation of the health assessment questionnaire II: A revised version of the health assessment questionnaire. Arthritis Rheum. 2004;50:3296–305.
5. Barber CEH, Zell J, Yazdany J, Davis AM, Cappelli L, Ehrlich-Jones L, et al. 2019 American College of Rheumatology Recommended Patient-Reported functional status assessment measures in rheumatoid arthritis. Arthritis Care Res (Hoboken). 2019;71:1531–9.
6. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: Dimensions and practical applications. Health Qual Life Outcomes. 2003;1:20.
7. Anderson J, Sayles H, Curtis JR, Wolfe F, Michaud K. Converting modified health assessment questionnaire (HAQ), multidimensional HAQ, and HAQII scores into original HAQ scores using models developed with a large cohort of rheumatoid arthritis patients. Arthritis Care Res (Hoboken). 2010;62:1481–8.
8. Taylor WJ, McPherson KM. Using Rasch analysis to compare the psychometric properties of the Short Form 36 physical function score and the Health Assessment Questionnaire disability index in patients with psoriatic arthritis and rheumatoid arthritis. Arthritis Rheum. 2007;57:723–9.
9. Oude Voshaar MA, Glas CA, ten Klooster PM, Taal E, Wolfe F, van de Laar MA. Crosscultural measurement equivalence of the Health Assessment Questionnaire II. Arthritis Care Res (Hoboken). 2013;65:1000–4.
10. Ten Klooster PM, Taal E, van de Laar MA. Rasch analysis of the Dutch Health Assessment Questionnaire disability index and the Health Assessment Questionnaire II in patients with rheumatoid arthritis. Arthritis Rheum. 2008;59:1721–8.
11. Waimann CA, Citera G-, del Pra FM, Marengo MF, Schneeberger EE, Sánchez M, et al. Validación de una versión argentina del Health Assessment Questionnaire-II (HAQ-II). Rev Arg Reumatol. 2011;22:21–9.
12. Haensch G. Español de América y español de Europa. Panace@. 2001;2:63–72.
13. Spanish Brief Survey Questionnaire (BSQ). Breve Encuesta Cuestionario [consultado 20 Feb 2018]. Disponible en: Https://www.arthritis-research.org/sites/default/files/documents/BSQSpanish.pdf.).
14. El español en los Estados Unidos: E pluribus unum? Estados Unidos de América: Ediciones ANLE; 2013.
15. Gonzalez VM, Stewart A, Ritter PL, Lorig K. Translation and validation of arthritis outcome measures into Spanish. Arthritis Rheum. 1995;38:1429–46.
16. Cella D, Hernandez L, Bonomi AE, Corona M, Vaquero M, Shiomoto G, et al. Spanish language translation and initial validation of the functional assessment of cancer therapy quality-of-life instrument. Med Care. 1998;36:1407–18.
17. Horta-Baas G, Romero-Figueroa MDS. Evaluacion de la intensidad del dolor en personas con artritis reumatoide mediante la escala de intensidad MOS. Med Clin (Barc). 2019;153:106–11.
18. Horta-Baas G. Reliability and validity of the Spanish version of the Medical Outcomes Study Pain Severity Scale in Mexican patient with rheumatic diseases. J Clin Rheumatol. 2020.
19. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO 3rd, et al. 2010 Rheumatoid arthritis classification criteria: An American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum. 2010;62:2569–81.
20. COSMIN Study design checklist for patient. Reported outcome measurement instruments. 2019 [consultado 30 Sep 2019]. Disponible en: Https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf.).
21. Pacheco-Tena C, Reyes-Cordero G, McKenna SP, Rios-Barrera VA. Adaptación y validación del *Rheumatoid Arthritis Quality of Life Scale* (RAQoL) al español de México. Reumatol Clin. 2011;7:98–103.
22. López-Alvarenga JC, Vázquez-Velázquez V, Arcila-Martínez D, Sierra-Ovando AE, González-Barranco J, Salín-Pascual RJ. [Accuracy and diagnostic utility of the Hospital Anxiety and Depression Scale (HAD) in a sample of obese Mexican patients] [artículo en español]. Rev Invest Clin. 2002;54:403–9.
23. García de Yébenes Prous MJ, Rodríguez-Salvanés F, Carmona-Ortells L. Validación de cuestionarios. Reumatol Clin. 2009;5:171–7.
24. Pérez-Gil JA, Chacón-Moscoso S, Moreno-Rodríguez R. Validez de constucto: El uso de análisis factorial exploratorio-confirmatorio para obtener evidencias de validez. Psicothema. 2000;12:442–6.

25. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine. A practical guide. Reino Unido: Cambridge University Press; 2011.
26. Brown A, Croudace TJ, Scoring. Estimating score precision using multimensional I.R.T. Models. En: Reise SP, Revicki DA, editores. Handbookof item response theory modeling. applications to typical performance assessment. Nueva York, EE. UU.: Routledge; 2015.
27. Attorresi HF, Lozzia GS, Abal FJP, Galibert MS, Aguerrí ME. Teoría de respuesta al ítem. Conceptos básicos y aplicaciones para la medición de constructos psicológicos. Rev Argent Clín Psicol. 2009;18:179–88.
28. Stover AM, McLeod LD, Langer MM, Chen WH, Reeve BB. State of the psychometric methods: Patient-reported outcome measure development and refinement using item response theory. J Patient Rep Outcomes. 2019;3:50.
29. Steichen TJ, Cox NJ. Concordance correlation coefficient. Stata J. 1998;43:35–9.
30. 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:S22–31.
31. Dinno A. Nonparametric pairwise multiple comparisons in independent groups using Dunnś test. Stata J. 2015;15:292–300.
32. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60:34–42.
33. Schneider L, Chalmers RP, Debelak R, Merkle EC. Model selection of nested and non-nested item response models using Vuong tests. Multivariate Behav Res. 2020;55:664–84.
34. Felt JM, Castaneda R, Tiemensma J, Depaoli S. Using person fit statistics to detect outliers in survey research. Front Psychol. 2017;8:863.
35. Byrne BM. Factor analytic models: Viewing the structure of an assessment instrument from three perspectives. J Pers Assess. 2005;85:17–32.
36. Calderón Garrido C, Navarro González D, Lorenzo Seva U, Ferrando Piera PJ. Multidimensional or essentially unidimensional? A multi-faceted factor-analytic approach for assessing the dimensionality of tests and items. Psicothema. 2019;31:450–7.
37. Clasificación internacional del funcionamiento, de la discapacidad y de la salud: CIF. Santander, España: Organización Mundial de la Salud; 2001.
38. Edwards MC, Wirth RJ, Houts CR, Bodine AJ. Three (or four) factors, four (or three) models. En: Reise SP, Revicki DA, editores. Handbook of item response theory modeling. Applications to typical performance assessment. Nueva York, EE. UU.: Routledge; 2015.
39. Reckase MD. Multidimensional item response theory. Nueva York, EE. UU.: Springer-Verlag Nueva York; 2009.