



Salud Pública de México

ISSN: 0036-3634

spm@insp.mx

Instituto Nacional de Salud Pública
México

Ramada-Rodilla, José María; Serra-Pujadas, Consol; Delclós-Clanchet, George L
Adaptación cultural y validación de cuestionarios de salud: revisión y recomendaciones
metodológicas

Salud Pública de México, vol. 55, núm. 1, enero-febrero, 2013, pp. 57-66

Instituto Nacional de Salud Pública
Cuernavaca, México

Disponible en: <https://www.redalyc.org/articulo.oa?id=10625594004>

- Cómo citar el artículo
- Número completo
- Más información del artículo
- Página de la revista en redalyc.org

redalyc.org

Sistema de Información Científica

Red de Revistas Científicas de América Latina, el Caribe, España y Portugal

Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

Adaptación cultural y validación de cuestionarios de salud: revisión y recomendaciones metodológicas

José María Ramada-Rodilla, MD, MOH,^(1,2) Consol Serra-Pujadas, MD, PhD,^(1,2,3)
George L Delclós-Clanchet, MD, MPH, PhD.^(2,3,4)

Ramada-Rodilla JM, Serra-Pujadas C,
Delclós-Clanchet GL.
Adaptación cultural y validación de cuestionarios de salud:
revisión y recomendaciones metodológicas.
Salud Publica Mex 2013;55:57-66.

Resumen

La traducción simple de un cuestionario puede dar lugar a interpretaciones erróneas debido a diferencias culturales y de lenguaje. Cuando se utilicen cuestionarios desarrollados en otros países e idiomas en estudios científicos, además de traducirlos, es necesaria su adaptación cultural y validación. El objetivo de este trabajo es revisar la literatura sobre la traducción, adaptación cultural y validación (TACV) de cuestionarios de salud, y sintetizar y proponer recomendaciones basadas en la literatura científica que faciliten este proceso. La TACV debe seguir un proceso sistematizado, por lo que se recomiendan dos etapas: a) adaptación cultural: traducción directa, síntesis, traducción inversa, consolidación por comité de expertos y pre-test, y b) validación (con hasta siete pasos): evaluación de la consistencia interna, fiabilidad intra e interobservador, validez lógica, de contenido, criterio y constructo. La falta de equivalencia de los cuestionarios limita las posibilidades de comparación entre poblaciones con idiomas o culturas diferentes y el intercambio de información en la comunidad científica.

Palabras clave: cuestionarios; escalas; encuestas de salud; comparación transultural; estudios de validación; confiabilidad y validez

Ramada-Rodilla JM, Serra-Pujadas C,
Delclós-Clanchet GL.
Cross-cultural adaptation and health questionnaires validation:
revision and methodological recommendations.
Salud Publica Mex 2013;55:57-66.

Abstract

The simple translation of a questionnaire may lead to misinterpretation due to language and cultural differences. When using questionnaires developed in other countries and languages in scientific studies it is necessary, besides the translation, to carry out a cross-cultural adaptation and validation. Our objective was to review the literature on cross-cultural adaptation and validation (CCAV) of health questionnaires, and to synthesize and propose recommendations based on the scientific literature to facilitate this process. The CCAV should follow a systematic process. Two steps are recommended: 1) cross-cultural adaptation: direct translation, synthesis, back translation, expert committee consolidation and pre-testing, and 2) validation (with up to seven steps): assessment of internal consistency, reliability, intra- and inter-observer reliability, face, content, criterion and construct validity. Lack of equivalence between questionnaires limits the comparability of results among populations with different cultures and languages and the exchange of information in the scientific community.

Key words: questionnaires; scales; health survey; cross-cultural comparison; validation studies; reliability and validity

- (1) Servicio de Salud Laboral, Parc de Salut MAR. Barcelona, España.
- (2) Centro de Investigación en Salud Laboral (CISAL), Universidad Pompeu Fabra. Barcelona, España.
- (3) CIBER de Epidemiología y Salud Pública (CIBERESP). Barcelona, España.
- (4) Epidemiology, Human Genetics and Environmental Sciences Division, The University of Texas School of Public Health. Houston, Texas, EUA.

Fecha de recibido: 2 de enero de 2012 • Fecha de aceptado: 21 de septiembre de 2012

Autor de correspondencia: José Ma. Ramada Rodilla. Centro de Investigación en Salud Laboral, Universidad Pompeu Fabra.

Dr. Aiguader, 88, 08003-Barcelona, España.

Correo electrónico: jramada@parcdesalutmar.cat

Valdría la pena imaginar a un investigador que se encuentra aplicando un cuestionario británico a una muestra de peatones alemanes. En el cuestionario se pregunta sobre la costumbre de “mirar a la derecha” antes de cruzar una vía de doble sentido de circulación. Es probable que se detecte una carencia en la formación vial de los peatones alemanes, ya que éstos no miran a la derecha cuando cruzan. Sin embargo, este resultado estará más bien relacionado con una inadecuada adaptación cultural del cuestionario porque en Alemania se circula por la derecha y, por tanto, “se mira a la izquierda” antes de cruzar.

La traducción simple de un cuestionario puede conducir a una interpretación errónea debido a diferencias culturales y de lenguaje. Si el proceso de traducción, adaptación cultural y validación (TACV) no se realiza correctamente pueden producirse errores de índole diversa dependiendo del objetivo del cuestionario.

Una inadecuada TACV de cuestionarios como el Goldberg (GHQ),¹ el Nordic Occupational Skin Questionnaire (NOSQ),² el Test de Control de Asma (ACT)³ o del Michigan Alcohol Screening Test (MAST),⁴ provocarían errores de clasificación en el despistaje de pacientes con trastornos ansioso-depresivos, dermatosis profesionales, asma o alcoholismo.

Deficiencias en la TACV de cuestionarios como el Work Ability Index (WAI),⁵ o el Work Role Functioning Questionnaire (WRFQ),⁶ podrían dar lugar a errores en la evaluación del grado de capacidad para el trabajo, afectando la orientación de las medidas preventivas.

Una TACV poco sistemática de cuestionarios para la vigilancia epidemiológica de enfermedades y exposiciones, como el Cuestionario de Detección Epidemiológica para Artritis Reumatoide,^{7,8} el Cuestionario Nórdico Estandarizado para la Detección de Síntomas Músculo-esqueléticos en Salud Ocupacional,⁹ o el Cuestionario para la Detección Integrada de Obesidad, Diabetes e Hipertensión Arterial de la Secretaría de Salud de México,¹⁰ podría llegar a inducir el diseño y puesta en marcha de políticas públicas inadecuadas.

La TACV es necesaria incluso cuando se desea aplicar un cuestionario en países distintos que hablan un mismo idioma. En ocasiones se asume que la adaptación cultural a un idioma diferente garantiza las propiedades psicométricas del cuestionario. Esto no siempre es así. Por ejemplo, las diferencias en cómo se realiza la actividad laboral en los países pueden modificar la validez de un cuestionario de aplicación en salud laboral.¹¹⁻¹³ La necesidad de intercambiar experiencias y llevar a cabo comparaciones entre poblaciones y países distintos precisa de versiones lingüísticas adecuadamente adaptadas y validadas de los instrumentos de medida.^{14,15}

El grado de cumplimiento de los pasos metodológicos que se recomiendan en la literatura internacional para la realización de la TACV es bajo. Para deobjetivar esta afirmación, se recuperaron todos los artículos, sin límite temporal ni de idioma, publicados en cinco de las revistas de epidemiología y salud pública con mayor factor de impacto en América Latina y en España -Revista Panamericana de Salud Pública, Revista de Saúde Pública, Salud Pública de México, Gaceta Sanitaria y Revista Española de Salud Pública-, usando los términos MeSH: cuestionarios, escalas, encuestas de salud, comparación transcultural, estudios de validación, confiabilidad y validez.

Se incluyeron aquellos artículos cuyo objetivo fue la TACV de un cuestionario a un idioma diferente del original. Se excluyeron aquellos que perseguían el diseño y validación de un cuestionario o bien la validación del mismo, partiendo de un cuestionario cuya adaptación cultural había sido publicada en un estudio anterior. Se obtuvieron en total 32 artículos que se analizaron en su versión completa. De ellos, 25% siguió menos de la mitad de los pasos recomendados; 72% siguió menos de 80% de dichos pasos, y sólo 6% de los artículos siguió la totalidad de éstos (cuadro I).

No se ha identificado ninguna revisión en la literatura que integre y sistematice todo el proceso de TACV, por lo que el objetivo de este trabajo fue revisar y sintetizar la literatura proponiendo recomendaciones que faciliten el proceso de TACV para su aplicación en cuestionarios de salud.

Material y métodos

Se realizó una revisión bibliográfica exhaustiva para localizar la información disponible sobre la metodología de la TACV de cuestionarios de salud.

La búsqueda bibliográfica se inició con la revisión de varios libros y monografías especializadas en metodología para el diseño, adaptación y validación de cuestionarios publicados entre 1996 y 2007.^{11,16-21} A partir de las citas bibliográficas de dichas publicaciones, se recuperaron diversos artículos sobre la TACV de cuestionarios de salud y sus aspectos metodológicos, que estuvieran publicados en inglés, francés, italiano, español y portugués. Se seleccionaron las palabras clave que agrupaban un mayor número de términos y se contrastaron con el tesoro de Medline, identificando los términos (MeSH terms): 1) “health survey”; 2) “health questionnaire”; 3) “scale”; 4) “cross cultural adaptation”; 5) “validation”; 6) “validity”, y 7) “reliability”. Con la combinación de estos términos se realizó la búsqueda en Medline, de tal manera que se obtuvieron 214 citas.

Cuadro I
CUMPLIMIENTO DE LOS PASOS METODOLÓGICOS PARA LA TACV DE CUESTIONARIOS PUBLICADOS EN LAS REVISTAS GS, RESP, SPM, RDSP, Y RPSP, SIN LÍMITE TEMPORAL NI DE IDIOMA, HASTA EL 1 DE NOVIEMBRE DE 2011. BARCELONA, ESPAÑA, NOVIEMBRE 2011

Artículo	Revista	Adaptación cultural				Validación								
		Traducción directa	Síntesis traducciones	Traducción inversa	Comité expertos	Pre-test	Fiabilidad		Validez					
							Consistencia interna	test-retest	Fiabilidad interobservador	Validez aparente	Validez contenido	Validez criterio	Validez constructo	
Mas Pons 1998	RESP	Sí	Sí	Sí	Sí	Sí	No	No	No	NP	No	No	No	No
López-Alvarenga 2001	SPM	Sí	No	No	No	No	No	No	Sí	NP	No	No	Sí	Sí
Amaral-Pinheiro 2002	RDSP	Sí	No	No	No	Sí	No	No	No	NP	Sí	Sí	Sí	No
Serra-Sutton 2002	RESP	Sí	Sí	Sí	Sí	Sí	No	No	No	NP	No	No	NP	No
López-Vázquez 2004	SPM	Sí	No	No	No	No	Sí	No	No	No	No	No	NP	Sí
Guimaraes de Mello 2004	RDSP	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	NP	No	No	No	No
Melgar-Quifonez 2005	SPM	NP	NP	NP	Sí	Sí	No	No	No	NP	Sí	Sí	Sí	No
Avanci 2005	RDSP	Sí	No	Sí	Sí	No	Sí	Sí	Sí	NP	Sí	Sí	No	Sí
Aymerich 2005	GS	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No	NP	Sí	Sí	No	Sí
Torres 2005	RDSP	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	NP	Sí	Sí	No	No
Majdalani 2005	RPSP	NP	NP	NP	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No
Rodriguez da Silva 2005	RDSP	Sí	No	Sí	Sí	Sí	Sí	Sí	Sí	No	Sí	No	Sí	Sí
López-Carmona 2006	SPM	NP	NP	NP	Sí	No	Sí	Sí	Sí	NP	No	No	Sí	No
Carpio 2006	RPSP	Sí	No	No	No	No	No	No	No	Sí	Sí	Sí	Sí	Sí
Álvarez 2006	SPM	NP	NP	NP	No	Sí	Sí	Sí	No	NP	No	No	Sí	Sí
Reichenheim 2007	RDSP	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	NP	Sí	Sí	Sí	Sí
Esteva 2007	GS	Sí	No	Sí	Sí	Sí	Sí	Sí	Sí	No	Sí	Sí	No	Sí
Pinto Guedes 2007	RDSP	Sí	Sí	Sí	Sí	Sí	No	No	Sí	NP	Sí	Sí	NP	Sí
Peña de León 2007	RPSP	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No	NP	No	No	Sí	Sí
Remor 2007	RDSP	Sí	Sí	Sí	No	No	Sí	Sí	No	NP	No	No	Sí	No
Aguirre Jaime 2008	RESP	NP	NP	NP	Sí	No	Sí	Sí	No	No	Sí	Sí	Sí	No
González-Block 2008	SPM	Sí	No	No	No	No	No	No	No	No	Sí	Sí	Sí	No
Pedro Gómez 2009	RESP	Sí	No	Sí	Sí	No	Sí	Sí	No	NP	No	No	NP	Sí
Zurbarán 2009	RPSP	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No	No	No	No	Sí	No
Martínez-Gómez 2009	RESP	Sí	No	Sí	Sí	Sí	Sí	Sí	Sí	NP	Sí	Sí	Sí	No
ShiNohara 2010	RDSP	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	NP	No	No	Sí	Sí
Silva 2010	RPSP	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	NP	Sí	Sí	No	No
de Souza-Machado 2010	RDSP	Sí	No	No	No	Sí	NP	NP	Sí	NP	No	No	No	No
Garrido-Urrutia 2010	RESP	Sí	Sí	No	No	No	Sí	Sí	Sí	NP	Sí	Sí	NP	No
Gutiérrez Sánchez 2011	RESP	Sí	No	Sí	No	No	Sí	Sí	No	NP	No	No	NP	Sí
Amaral Saliba 2011	RPSP	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	NP	Sí	Sí	Sí	Sí
de Barrios Leite 2011	RDSP	Sí	Sí	Sí	Sí	Sí	No	No	Sí	NP	No	Sí	No	Sí

TACV: Traducción, adaptación cultural y validación. GS: Gaceta Sanitaria. RESP: Revista Española de Salud Pública. SPM: Salud Pública de México
RDSP: Revista de Saúde Pública. RPSP: Revista Panamericana de Salud Pública. Sí: Paso realizado. No: Paso no realizado. NP: No procede

Fueron criterios de inclusión que el artículo tratara sobre aspectos metodológicos de los procesos de TACV de cuestionarios de salud y que fueran publicados en los idiomas mencionados. Con base en estos criterios y partiendo de la lectura de los resúmenes, se seleccionaron 20 artículos que se analizaron en su versión completa.^{12,13,15,22-38}

Asimismo, se realizó una búsqueda de la literatura gris a través de Internet, introduciendo como criterios de búsqueda las palabras clave obtenidas, así como los autores identificados en el proceso anterior. Finalmente, se incluyeron siete libros^{11,16-21} y 21 artículos.^{12-15,22-38} A partir de esta revisión, se elaboró una propuesta con las recomendaciones metodológicas sobre las que existía un mayor consenso entre los autores y se formuló un glosario con los términos más comúnmente empleados en los procesos de TACV de cuestionarios (cuadro II).

Síntesis y recomendaciones

Existe amplio consenso en recomendar dos etapas para el proceso de TACV: a) adaptación cultural, donde es necesario tener en cuenta los giros idiomáticos, el contexto cultural, y las diferencias en la percepción de la salud y la enfermedad de las poblaciones, y b) la validación en el idioma de destino, para evaluar el grado de preservación de las propiedades psicométricas.

Primera etapa: traducción y adaptación cultural

En esta etapa se traduce la herramienta partiendo de su versión original y procurando mantener la estructura del cuestionario. El objetivo es conseguir que el instrumento resultante mantenga la equivalencia semántica, idiomática, conceptual y experiencial con el cuestionario original.^{22,23} En la literatura existe consenso sobre cómo abordar esta primera etapa,^{12,13,22-27} recomendándose una secuencia de cinco pasos (figura 1):

Traducción directa: se realiza una traducción conceptual del instrumento. Deben participar, al menos, dos traductores bilingües independientes cuya lengua materna sea el idioma de destino. Uno de los traductores deberá conocer los objetivos y los conceptos considerados en el cuestionario, y tendrá experiencia previa en la traducción técnica de textos.

El otro u otros traductores no tendrán conocimientos previos sobre el cuestionario y desconocerán los objetivos del estudio. Estos traductores ofrecerán una traducción más ajustada al lenguaje de uso coloquial, detectando las dificultades de comprensión y traducción derivadas del uso de vocablos técnicos o poco comunes.

Cuadro II

GLOSARIO DE TÉRMINOS COMÚNMENTE EMPLEADOS EN LOS PROCESOS DE TRADUCCIÓN Y ADAPTACIÓN CULTURAL DE CUESTIONARIOS. BARCELONA, ESPAÑA, NOVIEMBRE 2011

- Adaptación cultural (cross-cultural adaptation):** tomar en consideración el contexto cultural, los giros idiomáticos y las diferencias en la percepción de la salud y la enfermedad de aquellas poblaciones en las cuales se desea aplicar.
- Consistencia interna (internal consistency reliability):** es el grado de interrelación y coherencia de los componentes (ítems o variables) del instrumento de medida.
- Constructo (construct):** teoría subyacente en el fenómeno o concepto que se quiere medir. Se trata de una cualidad no observable en una población de sujetos.
- Criterio o prueba de referencia (gold standard):** método de medición alternativo equivalente, independiente de los resultados de un cuestionario, fiable, exacto, objetivo y ampliamente aceptado como medida válida.
- Escala (scale):** graduación utilizada en diversos instrumentos de medida para posibilitar la medición de una magnitud.
- Especificidad (specificity):** capacidad para detectar a los individuos que no presentan el fenómeno de estudio.
- Fiabilidad (reliability):** grado en que un instrumento es capaz de medir sin errores. Es la proporción de la variancia total atribuible a diferencias verdaderas entre los sujetos.
- Fiabilidad inter-observador (inter-rater reliability):** mide el grado de acuerdo que hay entre dos o más evaluadores que valoran a los mismos sujetos con el mismo instrumento.
- Fiabilidad intra-observador o fiabilidad test-retest (test-retest reliability):** mide la estabilidad de las puntuaciones otorgadas por el mismo evaluador, en los mismos sujetos y con el mismo método en momentos diferentes.
- Ítem (item):** cada uno de los componentes o variables de un instrumento de medida; cada una de las partes o unidades de que se compone una prueba, un test o un cuestionario.
- Sensibilidad (responsiveness):** capacidad de detectar y medir cambios, tanto en los diferentes individuos como en la respuesta de un mismo individuo a través del tiempo.
- Traducción (translation):** expresar en una lengua algo que se ha expresado anteriormente o que está escrito en otra diferente.
- Traducción directa (forward translation):** es aquella que se realiza de un idioma extranjero al idioma del traductor.
- Traducción inversa (back translation):** es la traducción de un texto a su idioma original partiendo de una traducción de este texto realizada previamente a otro idioma.
- Traducción literal (literal translation):** es aquella en la que se respeta el sentido del texto original.
- Validación (validation):** evaluación del grado de preservación de las propiedades psicométricas del cuestionario.
- Validez (validity):** capacidad que tiene el instrumento de medir aquel constructo para el que ha sido diseñado.
- Validez aparente (face validity):** grado en que los ítems de un cuestionario, a juicio de los expertos y de los usuarios, miden de modo lógico o reflejan adecuadamente el constructo que se quiere medir.
- Validez de contenido (content validity):** grado en que el contenido de un instrumento es capaz de medir la mayor parte de las dimensiones del constructo que se quiere estudiar.
- Validez de constructo (construct validity):** grado en que las mediciones que resulten de las respuestas del cuestionario puedan considerarse como una medición del fenómeno estudiado.
- Validez de criterio (criterion validity):** grado en que el resultado del cuestionario predice o concuerda con algún criterio de "valor real" o gold standard.
- Valor predictivo positivo (positive predictive validity):** es la probabilidad de que esté presente el fenómeno de estudio en un individuo cuando el resultado del cuestionario es positivo.
- Valor predictivo negativo (negative predictive validity):** es la probabilidad de que no esté presente el fenómeno de estudio en un individuo cuando el resultado del cuestionario es negativo.

Todo el cuestionario, incluyendo las instrucciones, los ítems y las opciones de respuesta, se traducirá utilizando este método, recopilando todo en un informe.

Síntesis de traducciones: las traducciones serán comparadas por los traductores. Se identificarán y se discutirán las discrepancias entre las versiones traducidas hasta alcanzar el consenso. En el caso de que no exista consenso, se requerirá la participación del equipo de investigación. Al final, se realizará un informe del proceso en el que aparecerá una única traducción del cuestionario que será la versión de síntesis en el idioma de destino.

Traducción inversa (retro traducción): la versión de síntesis será retro traducida al idioma original, al menos por dos traductores profesionales bilingües cuya lengua materna sea la del cuestionario original. Los traductores trabajarán de forma independiente, estarán ciegos para la versión original del cuestionario, no tendrán conocimientos previos sobre el tema y desconocerán los objetivos del estudio.^{12,13}

Los traductores deberán subrayar las redacciones difíciles y las incertidumbres encontradas durante el

proceso de traducción. Se determinará si la traducción ha dado lugar a diferencias semánticas o conceptuales importantes entre el cuestionario original y la versión de síntesis obtenida en el paso anterior. Todo lo anterior se recopilará en un informe.

Consolidación por un comité de expertos: se recomienda constituir un comité multidisciplinar, si es posible de expertos bilingües en el tema sobre el que trata el cuestionario: un experto en metodología, un lingüista y un profesional de la salud, además de los traductores que han participado en el proceso. El objetivo de este comité será llegar a un único cuestionario consolidado pre-final adaptado al idioma de destino.^{16,17}

En este paso se dispondrá de las traducciones directas (paso 1), la versión de síntesis (paso 2) y las retrotraducciones (paso 3). Se identificarán y discutirán las discrepancias encontradas. Se cerciorará de que la versión pre-final sea totalmente comprensible y equivalente al cuestionario original. Se asegurará que el cuestionario pre-final resulte comprensible para una persona escolarizada con conocimientos equivalentes a un individuo de 12 años de edad.

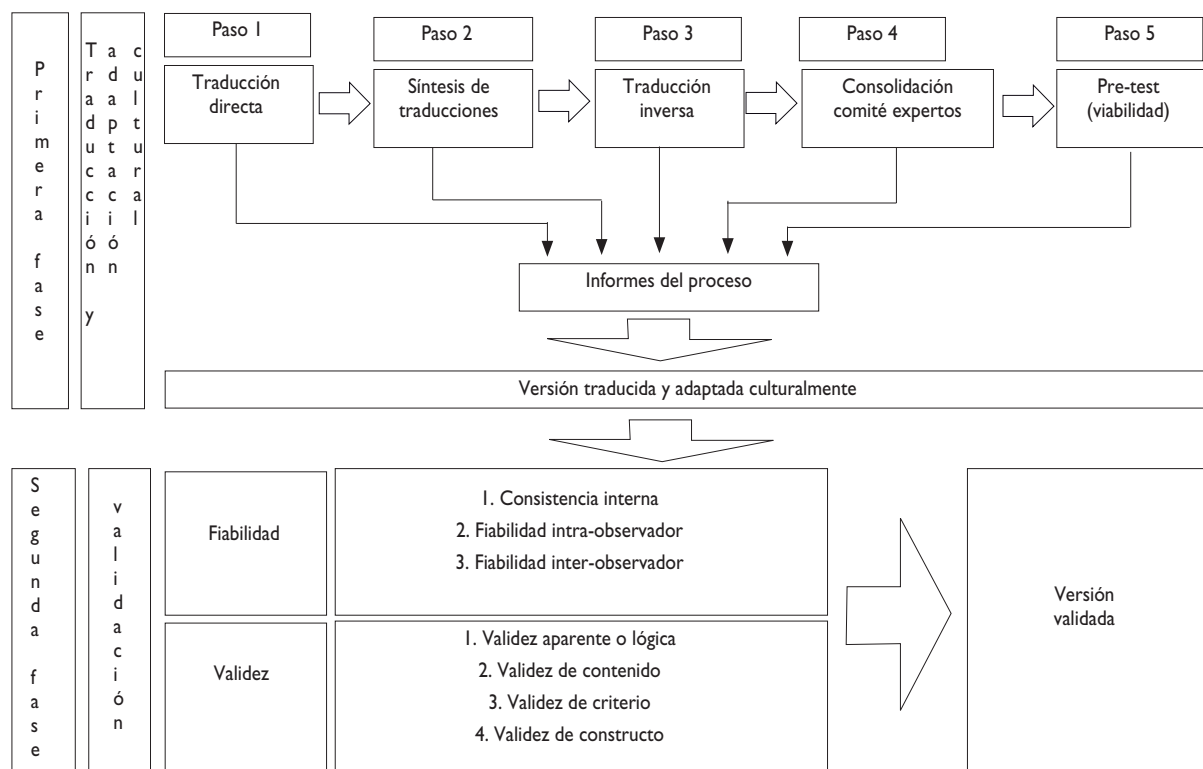


FIGURA 1. PROCESO DE TRADUCCIÓN, ADAPTACIÓN CULTURAL Y VALIDACIÓN (ADAPTADO DE REFERENCIA 22). BARCELONA, NOVIEMBRE 2011

En el caso de que surjan incertidumbres se recurrirá, de ser posible, con alguno de los autores del cuestionario para solicitar su participación. Se elaborará un informe que sintetice las decisiones del comité, incluyendo la versión consolidada.

Pre-test (aplicabilidad / viabilidad): su realización permitirá evaluar la calidad de la traducción, la adaptación cultural y la aplicabilidad o viabilidad del cuestionario. Asimismo permitirá calcular si el tiempo de cumplimentación se encuentra dentro de límites razonables.

Investigadores como Durand y colaboradores,²⁵ y Gallasch y colaboradores,²⁶ realizaron el pre-test durante el proceso de traducción y adaptación cultural del Work Role Functioning Questionnaire (WRFQ-27) con una muestra de 30-40 trabajadores, y se obtuvieron resultados satisfactorios. Lo mismo realizaron De Soárez y colaboradores para el Work Limitations Questionnaire (WLQ), incluyendo a 20 voluntarios.²⁷ Beaton propuso incluir en la muestra entre 30 y 40 participantes, basándose en una revisión bibliográfica de adaptaciones culturales.²² Se recomienda la realización del pre-test con participantes de distintos niveles educativos y, si se trata de cuestionarios autocumplimentados, los participantes deberán saber leer y comprender lo leído.

Para seleccionar la muestra, es importante definir los criterios de inclusión y exclusión, así como el modo en que serán reclutados los participantes. En el caso de cuestionarios de aplicación en salud laboral, se recomienda incluir en el pre-test a trabajadores en activo, con una jornada mayor o igual a 10 horas semanales, de ambos sexos, con edades entre 18 y 65 años, con diferentes niveles educativos y que hablen como primera lengua, lean y comprendan el idioma de destino si se trata de cuestionarios autocumplimentados. De cada participante se recopilarán datos, al menos, sobre sus características sociodemográficas, nivel educativo y ocupación.^{25,26}

Se solicitará a los participantes que llenen la versión consolidada y, mediante una entrevista estructurada, se les invitará a comentar cualquier aspecto que haya resultado difícil de entender. Se recomienda grabar estas entrevistas así como la autorización previa de los participantes, con el fin de poder revisarlas tantas veces como sea necesario.

Al final, se realizará un informe donde se identificarán las posibles dificultades en la comprensión de las instrucciones del cuestionario, las preguntas y las opciones de respuesta. Se recomienda la revisión de cualquier pregunta del cuestionario si al menos 15% de los participantes encuentran dificultades en la misma.²⁷

Segunda etapa: validación del cuestionario en el idioma destino

La correcta traducción y adaptación cultural de un cuestionario no siempre garantiza la preservación de sus propiedades psicométricas, por lo que es necesaria su validación en el idioma de destino.²²

Para que un cuestionario se considere válido, debe reunir las siguientes características: a) ser fiable y capaz de medir sin error; b) ser capaz de detectar y medir cambios, tanto entre individuos como en la respuesta de un mismo individuo a través del tiempo; c) ser sencillo, viable y aceptado por pacientes, usuarios e investigadores; d) ser adecuado para medir el fenómeno que se pretende medir, y e) reflejar la teoría subyacente en el fenómeno o concepto que se quiere medir. Todas estas características están relacionadas con dos propiedades de los cuestionarios: la fiabilidad y la validez.¹⁴

La Sociedad Internacional para la Evaluación de la Calidad de Vida (en inglés, IQOLA)^{8,18,19} y otros investigadores como Aday,¹⁹ Lam,³⁰ Mokkink,³¹⁻³³ Ren,³⁴ Scott-Lennox³⁵ y Wiesinger,³⁶ han propuesto o empleado diferentes métodos de evaluación de la fiabilidad y validez de los cuestionarios. De acuerdo con esas experiencias, se propone la validación de cuestionarios con la siguiente secuencia (figura 1):

1. *Fiabilidad:* es el grado en que un instrumento es capaz de medir sin errores. Mide la proporción de variación en las mediciones que es debida a la diversidad de valores que adopta la variable y no al posible error sistemático o aleatorio.^{14,33} La fiabilidad determina la proporción de la variancia total atribuible a diferencias verdaderas entre los sujetos.^{20,33,37}

Dependiendo de las características del cuestionario, su fiabilidad puede evaluarse para todas o algunas de sus tres dimensiones: 1) consistencia interna; 2) fiabilidad intra-observador o fiabilidad test-retest, y 3) fiabilidad inter-observador.

- 1.1. *Consistencia interna:* es el grado de interrelación y coherencia de los ítems. A través de este aspecto, se evalúa si los ítems que miden un mismo constructo presentan homogeneidad entre ellos.^{33,39} Cuando la escala de un instrumento es consistente, se garantiza que todos los ítems miden un solo constructo y, en general, se asegura la existencia de una relación lineal entre la suma de las puntuaciones de los ítems y el constructo medido.

Un constructo es una cualidad latente o intangible de un sujeto o de una población que no se puede observar y medir directamente con un instrumento de medida, ya que esta cualidad tiene lugar dentro de una teoría. Son ejemplos el estrés laboral, la motivación, la discapacidad o el liderazgo. Evaluar la fiabilidad de un instrumento no ofrece mayores problemas cuando se trata de cuantificar cualidades objetivas, como el peso o la talla. No obstante, para los constructos es necesario probar de forma empírica que el instrumento sirve para medir aquello que se pretende medir.

La medición de los constructos se realiza frecuentemente mediante cuestionarios donde se supone que cada ítem está relacionado con la cualidad no observable de interés. Para cada ítem se suele solicitar una respuesta a la que se asigna una puntuación. La suma de las puntuaciones proporciona la escala del cuestionario. En ocasiones, una escala puede estar compuesta por un grupo de subescalas. Por ejemplo, el riesgo laboral psicosocial es un constructo que, a su vez, puede estar compuesto por varias dimensiones como el nivel de demanda del trabajo, las recompensas, el nivel de control y el apoyo social.

El coeficiente alfa de Cronbach permite cuantificar el nivel de fiabilidad de una escala si se cumplen dos requisitos: a) debe estar formada por un conjunto de ítems, cuyas puntuaciones se suman para calcular una puntuación global, y b) todas las puntuaciones de los ítems deben medir en la misma dirección; por ejemplo, a mayor puntuación mayor capacidad funcional o mayor bienestar emocional.

El coeficiente alfa de Cronbach es la media ponderada de las correlaciones entre los ítems que forman parte de una escala.³⁹ Cuando el instrumento está compuesto por un grupo de subescalas, debe calcularse el coeficiente alfa de Cronbach para los ítems respecto de la puntuación global (correlación ítem-total) y para los ítems de cada subescala respecto del valor de la misma (correlación ítem-subescala).

El coeficiente alfa de Cronbach no viene acompañado de ningún valor de p que permita rechazar o no la hipótesis de fiabilidad de la escala. Puede adoptar valores entre 0 y 1. Se considera que valores alfa superiores a 0.70 son suficientes para garantizar la consistencia interna de la escala.

1.2 *Fiabilidad intra-observador o fiabilidad test-retest*: este aspecto hace referencia a la repetibilidad del instrumento, cuando se administra con el mismo método a la misma población en dos momentos diferentes.^{14,33} Cuando la escala es cuantitativa, su análisis se realiza mediante el cálculo del coeficiente de correlación intraclase (CCI), y cuando es cualitativa se realiza mediante el cálculo del índice Kappa de Cohen.^{21,37}

El tiempo que debe transcurrir entre la primera vez (test) y la segunda (retest) dependerá de lo que se esté midiendo. No debe ser muy largo para evitar que el fenómeno observado sufra variaciones que alterarían el valor de la repetibilidad y tampoco debe ser demasiado corto para evitar el recuerdo de las respuestas (efecto aprendizaje).

1.3 *Fiabilidad inter-observador*: es el grado de acuerdo que hay entre dos o más evaluadores que valoran a los mismos sujetos con el mismo instrumento.³³ Esta propiedad no es evaluable cuando se trata de cuestionarios autocumplimentados, ya que es el propio individuo quien proporciona las respuestas sin que exista interferencia de los investigadores.

Si se requiere su evaluación, se realizará mediante el cálculo del coeficiente de correlación intraclase (CCI) cuando la escala sea cuantitativa, y el índice Kappa de Cohen cuando sea cualitativa. Las limitaciones principales se deben a la posibilidad de que existan de acuerdos entre los observadores debidos al azar y la posibilidad de que exista un error sistemático (sesgo de información) de alguno de los evaluadores.

2. *Validez*: es la capacidad del cuestionario de medir aquel constructo para el que ha sido diseñado.^{19,33} Puede evaluarse para todas o sólo para alguna de sus cuatro dimensiones: validez aparente o lógica, de contenido, de criterio y de constructo.

2.1 *Validez aparente o lógica*: se refiere al grado en que un cuestionario, a juicio de los expertos y de los usuarios, mide de forma lógica lo que quiere medir.^{14,19}

Cuando se carece de validez aparente o lógica, los sujetos sometidos a estudio pueden no ver la relación entre las preguntas que se les formulan y el objeto para el cual han accedido a contestar. Este hecho puede provocar el rechazo de los participantes.

Esta dimensión de la validez debe evaluarse en el momento de su diseño; no obstante, si en el proceso de TACV se detectan desajustes debidos al proceso de traducción o adaptación cultural, será necesario corregirlos.

- 2.2 *Validez de contenido*: los constructos suelen estar compuestos por varias dimensiones. La validez de contenido es el grado en que la herramienta es capaz de medir la mayor parte de las dimensiones del constructo.^{14,19,33} Un cuestionario con alta validez de contenido es aquel que mide todas las dimensiones relacionadas con el constructo que se quiere estudiar.

Su evaluación es un proceso formal que siempre debe realizarse en un proceso de TACV y consiste en valorar si los ítems del cuestionario son una muestra representativa de aquello que se quiere medir. Se trata de una evaluación empírica, basada en juicios de diferente procedencia, como son las opiniones de los autores de la herramienta, los resultados de estudios piloto, los razonamientos realizados por el comité de expertos en un proceso de TACV y el análisis cualitativo de los comentarios realizados por los participantes durante el proceso de pre-test.

- 2.3 *Validez de criterio*: establece la validez de un instrumento comparándola con algún criterio externo o prueba de referencia ("gold standard", GS). Tiene dos dimensiones: 1) la validez concurrente o grado en que el resultado del cuestionario concuerda con algún GS, y 2) la validez predictiva o grado en que es capaz de pronosticar un determinado resultado.^{14,19,33}

El GS debe ser un método alternativo equivalente, independiente de los resultados del cuestionario, fiable, exacto, objetivo y ampliamente aceptado como medida válida.^{14,19} Cuando reúne estos requisitos es capaz de dar un resultado siempre positivo en presencia del fenómeno a estudiar y siempre negativo en ausencia del mismo. Por ejemplo, la electromiografía realizada en condiciones adecuadas podría ser el GS frente a un cuestionario para la evaluación de la presencia del síndrome del túnel carpiano.

Siempre que haya un GS, debería evaluarse la validez de criterio concurrente, siguiendo cinco pasos: 1) selección del GS; 2) selección de una muestra de sujetos representativa de la población; 3) administración del cuestionario y obtención del resultado para cada individuo; 4) evaluación de cada individuo con el GE, y 5)

comparación de los resultados obtenidos con el cuestionario y el GS.

El análisis de la validez de criterio concurrente consiste en examinar la fuerza de la correlación existente entre el resultado del cuestionario y el del GS y se puede cuantificar mediante el cálculo del coeficiente de correlación de Pearson (r). Otro enfoque para cuantificar la validez de criterio concurrente consiste en analizar la sensibilidad y la especificidad.^{19,21}

La sensibilidad es la capacidad que tiene el cuestionario para detectar a los individuos que presentan el fenómeno de estudio. Se puede definir como la probabilidad de que un individuo que realmente tenga el fenómeno de estudio obtenga un resultado positivo cuando se le aplique el cuestionario. Se calcula mediante el cociente entre los verdaderos positivos (VP) y la suma de los VP y los falsos negativos (FN). De ahí que también que se le conozca como la fracción de verdaderos positivos (FVP). $\text{Sensibilidad} = \text{VP} / (\text{VP} + \text{FN})$.

La especificidad es la capacidad de detectar a los que no presentan el fenómeno de estudio, y es la probabilidad de que un individuo que no tenga el fenómeno de estudio obtenga un resultado negativo cuando se le aplique el cuestionario. Se puede calcular mediante el cociente entre los VN y la suma de los VN y los FP, y se le conoce como la fracción de verdaderos negativos (FVN); $\text{especificidad} = \text{VN} / (\text{VN} + \text{FP})$ (cuadro III).

Cuanto más alta sea la sensibilidad y especificidad, y menor sea el porcentaje de FP y FN, mayor será la validez concurrente.

Cuadro III
CÁLCULO DE LA SENSIBILIDAD, ESPECIFICIDAD, VALOR PREDICTIVO POSITIVO Y VALOR PREDICTIVO NEGATIVO.*
BARCELONA, ESPAÑA, NOVIEMBRE 2011.

Resultado del cuestionario	Fenómeno de estudio (gold standard)		
	Presente	Ausente	Total
Positivo	VP	FP	VP+FP
Negativo	FN	VN	FN+VN
Total	VP+FN	FP+VN	

Fuente: * Adaptado de referencia 21

VP: verdaderos positivos; FP: falsos positivos; FN: falsos negativos; VN: verdaderos negativos

Sensibilidad: $\text{VP} / (\text{VP} + \text{FN})$

Especificidad: $\text{VN} / (\text{FP} + \text{VN})$

Valor Predictivo Positivo (VPP): $\text{VP} / (\text{VP} + \text{FP})$

Se considera que un cuestionario tiene una sensibilidad y especificidad aceptable cuando éstas son superiores a 0,80.²⁰ A partir de aquí, puede ser de interés conocer la validez predictiva.²¹ El valor predictivo positivo (VPP) es la probabilidad de que un individuo presente el fenómeno de estudio que se busca medir con el cuestionario si se obtiene un resultado positivo en el mismo. Se calcula mediante la proporción de participantes con un resultado positivo en el cuestionario y que finalmente presentaban el fenómeno de estudio que se intentaba medir: $VPP = VP / (VP + FP)$. El valor predictivo negativo (VPN) es la probabilidad de que no esté presente dicho fenómeno cuando el resultado del cuestionario es negativo: $VPN = VN / (FN + VN)$.

- 2.4 *Validez de constructo*: es el grado en que las mediciones que resultan de las respuestas del cuestionario pueden considerarse una medición del fenómeno estudiado.^{14,19,33} Su evaluación consiste en contrastar las hipótesis que se han formulado sobre el comportamiento de las puntuaciones de un instrumento en situaciones diferentes.

Existen varios métodos para su evaluación, que deben realizarse cuando el fenómeno a medir es abstracto o no es posible comparar con un GE. El uso de técnicas de análisis de la validez para grupos conocidos es un procedimiento muy adecuado en cuestionarios de salud laboral para medir el grado de capacidad física o cognitiva para el trabajo. Permite comparar los resultados obtenidos mediante la aplicación del cuestionario a grupos con un diagnóstico clínico conocido de salud física o mental.^{19,20}

Conclusiones

La TACV de cuestionarios para su uso en otros idiomas es un proceso que consume recursos; sin embargo, cuando se lleva a cabo de forma sistemática permite obtener una herramienta de medición equivalente a su versión original.

El modo en que se realiza la TACV de cuestionarios de salud es perfectible; así entonces, es importante seguir las recomendaciones metodológicas. Si el proceso de TACV no se lleva a cabo de manera rigurosa, pueden producirse errores con implicaciones en el diagnóstico, en las decisiones que deben tomarse con respecto a la terapia individual, en los registros epidemiológicos e, incluso, en el diseño y puesta en marcha de políticas públicas. Además, el uso de herramientas no equivalentes

al cuestionario original puede producir resultados no fiables o confusos que podrían limitar el intercambio de información entre la comunidad científica.^{13,14,22-24}

Esta propuesta para la TACV de cuestionarios de salud guarda coherencia con las recomendaciones de expertos como Alexandre,¹³ Beaton,²² Carvajal,²³ Guillemin¹² y Herdman²⁴ para la realización de traducciones y adaptaciones culturales.

El proceso de traducción y adaptación debe ir seguido de un proceso de validación en la lengua de destino, lo cual permite minimizar el sesgo de información que podría asociarse a la administración de cuestionarios en países con idiomas y culturas diferentes. Por ello, se complementa el proceso proponiendo una serie de pasos a seguir durante la etapa de validación, coherentes con las recomendaciones de expertos como Aday,¹⁹ Mokkink,³¹⁻³³ Müller³⁷ y Keszei.³⁸

Declaración de conflicto de intereses: Los autores declararon no tener conflicto de intereses.

Referencias

- Goldberg D, Bridges K, Duncan-Jones P, Grayson D. Detecting anxiety and depression in general medical settings. *BMJ* 1988; 297: 897-899.
- Susitaival P, Flyvholm MA, Meding B, Kanerva L, Lindberg M, Svensson A, et al. Nordic Occupational Skin Questionnaire (NOSQ-2002): a new tool for surveying occupational skin diseases and exposure. *Contact Dermatitis* 2003;49:70-76.
- Melosini L, Dente FL, Bacci E, Bartoli ML, Cianchetti S, Costa F, et al. Asthma control test (ACT): comparison with clinical, functional, and biological markers of asthma control. *J Asthma* 2012;49:317-323.
- Connor JP, Grier M, Feeney GF, Young RM. The validity of the Brief Michigan Alcohol Screening Test (bMAST) as a problem drinking severity measure. *J Stud Alcohol Drugs* 2007;68:771-779.
- Tuomi K, Ilmarinen J, Eskelinen L, Järvinen E, Toikkanen J, Klockars M. Prevalence and incidence rates of diseases and work ability in different work categories of municipal occupations. *Scand J Work Environ Health* 1991;17 (Suppl 1):67-74.
- Amick BC III, Lerner D, Rogers WH, Rooney T, Katz JN. A review of health-related work outcome measures and their uses, and recommended measures. *Spine* 2000; 25:3152-160.
- Scublinksky D, González C, Iannantuono R, Somma LF, Rillo O, Casado G et al. Adaptación al español y validación del cuestionario de detección epidemiológica para artritis reumatoidea. *Rev Argent Reumatol* 2008; 19:33-35.
- Simonsson M, Bergman S, Jacobsson L, Petersson I, Svensson B. The prevalence of rheumatoid arthritis in Sweden. *Scand J Rheumatol* 1999;28:340-343.
- Kuorinka I, Jonsson B, Kilbom A, Vinterberg H, Biering-Sørensen F, Andersson G, et al. Standardised Nordic questionnaires for the analysis of musculoskeletal symptoms. *Appl Ergon* 1987; 18: 233-237.
- Tapia-Conyer R, Velázquez-Monroy O, Lara-Esqueda A, Tapia-Olarte F, Aurora-Jiménez R, Sánchez-Montes J, et al. Guía de detección integrada de obesidad, diabetes e hipertensión arterial. [monografía en Internet]. Ciudad de México, DF: Secretaría de Salud de México; [consultado 2012 septiembre 18]. Disponible en: www.salud.gob.mx/unidades/cdi/documentos/DOCSAL7482.pdf

11. Hutchinson A, Bentzen N, Konig-Zahn C. Cross cultural health outcome assessment: a user's guide. The Netherlands: ERGHO, 1996.
12. Guillemin F. Cross-cultural adaptation and validation of health status measures. *Scand J Rheumatol* 1995;24:61-63.
13. Alexandre NMC, Guirardello Ede B. Cultural adaptation of instruments utilized in occupational health. *Rev Panam Salud Publica* 2002;11:109-111.
14. García de Yébenes MJ, Rodríguez-Salvanés F, Carmona-Ortells L. Validación de cuestionarios. *Reumatol Clin* 2009;5:171-177.
15. Kulis D, Arnott M, Greimel ER, Bottomley A, Koller M. Trends in translation requests and arising issues regarding cultural adaptation. *Expert Rev Pharmacoecon Outcomes Res* 2011;11:307-314.
16. Lobiondo-Wood G, Haber J. Reliability and validity. *Nursing research: methods, critical appraisal, and utilization*. 4a. ed. St. Louis: Mosby, 1998.
17. Burns N, Grove SK. The practice of nursing research: conduct, critique and utilization. 3a. ed. Philadelphia: Saunders, 1997.
18. Ware JE Jr, Gandek B, Keller S, IQOLA Group. Evaluating instruments used cross-nationally: Methods from the IQOLA Project. En: SpilkerB, ed. *Quality of life and pharmacoeconomics in clinical trials*. 2a. ed. Philadelphia: Lippincott-Raven Publishers, 1996: 681-692.
19. Aday LA, Cornelius LJ. Designing and conducting health surveys: a comprehensive guide. 3a. ed. San Francisco, CA: Jossey-Bass publisher, 2006.
20. Argimon-Pallas JM, Jimenez-Villa J. Métodos de investigación clínica y epidemiológica. 3a. ed. Madrid: Elsevier España, 2004.
21. Serra C, Company A. Vigilancia de la salud. En: Ruiz-Frutos C, García AM, Delclòs J, Benavides FG. *Salud laboral, conceptos y técnicas para la prevención de riesgos laborales*. 3a. ed. Barcelona: Masson, 2007: 255-264.
22. Beaton DE, Bombardier C, Guillemin F, Bosi-Ferraz M. Guidelines for the process of cross-cultural adaptation of self-reports measures. *Spine* 2000;25:3186-3191.
23. Carvajal A, Centeno C, Watson R, Martínez M, Rubiales AS. How is an instrument for measuring health to be validated?. *An Sist Sanit Navar* 2011;34:63-72.
24. Herdman M, Fox-Rushby J, Badia X. A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. *Qual Life Res* 1998;7:323-335.
25. Durand MJ, Vachon B, Hong QN, Imbeau D, Amick BC III, Loiselle P. The cross-cultural adaptation of the work role functioning questionnaire in Canadian French. *Int J Rehabil* 2004;27:261-268.
26. Gallasch CH, Alexandre NM, Amick B 3rd. Cross-cultural adaptation, reliability, and validity of the work role functioning questionnaire to Brazilian Portuguese. *J Occup Rehabil* 2007;17:701-711.
27. de Soárez PC, Kowalski CC, Ferraz MB, Ciconelli RM. Translation into Brazilian Portuguese and validation of the Work Limitations Questionnaire. *Rev Panam Salud Publica* 2007;22:21-28.
28. Bullinger M, Aonso J, Apolone G, et al. Translating health status questionnaires and evaluating their quality: the IQOLA Project approach. *International Quality of Life Assessment*. *J Clin Epidemiol* 1998;51:913-923.
29. Gandek B, Ware JE Jr, IQOLA Group. Methods for validation and norming translations of health status questionnaires: the IQOLA project approach. *International quality of life assessment*. *J Clin Epidemiol* 1998;51:953-959.
30. Lam CL, Gandek B, Ren XS, Chan MS. Tests of scaling assumptions and construct validity of the Chinese (HK) version of the SF-36 Health Survey. *J Clin Epidemiol* 1998;51:1139-1147.
31. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* 2010;19:539-549.
32. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Med Res Methodol* 2010;10: 22.
33. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;63:737-745.
34. Ren XS, Amik B III, Zhou L, Gandek B. Translation and psychometric evaluation of a Chinese version of the SF-36 Health Survey in the United States. *J Clin Epidemiol* 1998;51:1129-1138.
35. Scott-Lennox JA, Wu AW, Boyer JG, Ware JE Jr. Reliability and validity of French, German, Italian, Dutch, and UK English translations of the medical outcomes study HIV Health Survey. *Med Care* 1999;37:908-925.
36. Wiesinger GF, Nhur M, Quitann M, Ebenbichler G, Wölfl G, Fialka-Moser V. Cross-cultural adaptation of the Roland-Morris questionnaire for German-speaking patients with low back pain. *Spine* 1999;24:1099-1103.
37. Müller R, Büttner P. A critical discussion of intraclass correlation coefficients. *Stat Med* 1994;13:2465-2476.
38. Keszei AP, Novak M, Streiner DL. Introduction to health measurement scales. *J Psychosom Res* 2010;68:319-323.
39. Cronbach, LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16:297-334.