Criterios de cumplimiento de las directrices de la ITC para la adaptación de test

A. Hernández[1], M. D. Hidalgo[2], R. K. Hambleton[3] and J. Gómez-Benito[4].

[1] Universitat de València, [2] Universidad de Murcia, [3] Universidad de Massachusetts en Amherst, y [4] Universitat de Barcelona

Los criterios de cumplimiento de las directrices de la ITC (International Test Comission) que se presentan en este documento se complementan con el artículo que describe el proceso seguido para su desarrollo. Para hacer referencia a los criterios de cumplimiento propuestos en este documento, se debe utilizar la referencia correspondiente a dicho artículo:

Hernández, A., Hidalgo, M. D., Hambleton, R. K. & Gómez-Benito, J. (2020). International Test Commission guidelines for test adaptation: A criterion checklist. *Psicothema, 32* (3), 390-398. doi: 10.7334/psicothema2019.306

Introducción

La segunda edición de las directrices de la ITC para la traducción y adaptación de test (ITC, 2017) puede descargarse en: https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

Una versión previa de esta segunda edición se publicó en español en 2013 (Muñiz, Elosua y Hambleton, 2013). En dicha edición aparecen un total de 19 directrices, en vez de las 18 finales de 2017. Cuando las directrices coinciden, se presenta la traducción literal utilizada por Muñiz et al. (2013; pp. 153-154). En este caso las directrices aparecen entrecomilladas.

De acuerdo con las directrices de la ITC, las expresiones 'test' y 'testing' deben interpretarse con un significado amplio. Por lo tanto, en los criterios de cumplimiento que proponemos, el término 'test' se refiere a todo tipo de test (psicológicos, como los de inteligencia y personalidad; educativos y de recursos humanos, incluyendo las baterías de test cognitivos y de evaluación de competencias, las pruebas clínicas, etc.) y

abarca distintos tipos de formato (cuestionarios, escalas comportamentales, de evaluación de rendimiento, y otras herramientas de evaluación).

Las 18 directrices para la adaptación de test propuestas por la ITC (2017, segunda edición) se organizan en 6 categorías amplias (directrices preliminares, de desarrollo, de confirmación, de administración, de puntuación e interpretación y, finalmente, de documentación). En este documento, las 18 directrices se han operacionalizado a través de una serie de criterios (29 en total). Se han considerado todas y cada una de las directrices propuestas por la ITC. Sin embargo, en algunas ocasiones, hemos invertido el orden de presentación de las directrices, con el fin de facilitar la evaluación del grado de cumplimiento de los criterios propuestos. Cada uno de los criterios tiene una etiqueta alfanumérica que identifica la categoría a la que corresponde la directriz y el criterio específico dentro de dicha categoría. La etiqueta también incluye, entre paréntesis, dos contadores. El primero se refiere a la directriz y el segundo al criterio propuesto por nosotros, referido a dicha directriz. Por ejemplo, DP3-1 (D3, C4) se refiere al primer criterio de la tercera directriz en la categoría de "Directrices Previas" (DP). Asimismo, se trata de la tercera directriz propuesta por la ITC, y el cuarto criterio propuesto por nosotros De igual modo, DD3-2 (D6, C9) se refiere al segundo criterio propuesto para operacionalizar la tercera directriz de la ITC sobre "Directrices de Desarrollo" (DD), que corresponde a la sexta directriz de la ITC, y al noveno criterio de nuestra propuesta de operacionalización

1. DIRECTRICES PREVIAS

Estas directrices se refieren a la planificación de la adaptación, considerando una serie de aspectos preliminares.

DP1. "Antes de comenzar con la adaptación hay que obtener los permisos pertinentes de quien ostente los derechos de propiedad intelectual del test."

DP2: Evaluar si el grado de solapamiento en la definición y contenido del constructo medido mediante el test, así como en el contenido de los ítems, es suficiente para el uso (o usos) previsto(s) de las puntuaciones en las poblaciones de interés.

DP3. Minimizar la influencia de cualquier diferencia cultural o lingüística, que sea irrelevante para el uso previsto del test en las poblaciones de interés.

Operacionalización

Antes de considerar la directriz DP1, referida a los permisos y derechos de autor, es necesario determinar si, teniendo en cuenta la directriz DP2, el grado de solapamiento en el constructo entre las poblaciones de interés es suficiente para aconsejar la traducción/adaptación del test. Por consiguiente, comenzamos con la directriz DP2 para pasar luego a las directrices DP1 y DP3.

DP2-1 (G2, C1): Proporcionar evidencia teórica y empírica de que el constructo de interés es relevante para la población diana a la que va dirigida la versión adaptada (van der Vijver y Leung, 2011).

Excelente: Hay razones teóricas (basadas en artículos, juicios de expertos, etc.) y evidencia empírica (por ejemplo, obtenida mediante entrevistas, observaciones o encuestas a individuos de la población diana) que apoyan que el constructo es relevante para la población diana

Aceptable: Hay razones teóricas (basadas en artículos, juicios de expertos, etc.) que sugieren que el constructo es relevante para la población diana. Sin embargo, no se ha recogido evidencia empírica sobre la relevancia del constructo.

DP2-2 (G2, C2): Considerar si el significado del constructo puede generalizarse a través de las culturas, y justificar que la traducción/adaptación del test es preferible a la creación de un test nuevo dirigido a la población diana (Hambleton, Merenda, y Spielberg, 2005; van der Vijver y Leung, 2011).

Excelente: Un grupo de expertos en el constructo y en las culturas y lenguas implicadas evalúa la definición del constructo, su dimensionalidad, y si los ítems del test original captan adecuadamente la definición y dimensionalidad del constructo. Tras dicha evaluación justifican que a) hay un solapamiento completo del constructo en las poblaciones, y b) los ítems originales son adecuados para representar el constructo en la población diana.

Aceptable: Un grupo de expertos en el constructo y en las culturas y lenguas implicadas, evalúa la definición del constructo, su dimensionalidad, y si los ítems del test original captan adecuadamente la definición y dimensionalidad del constructo. Tras dicha evaluación concluyen que existe un solapamiento parcial del constructo en las poblaciones y que un número significativo de los ítems originales son adecuados para representar el constructo en la población diana.

DP1-1 (G1, C3)*: Si la adaptación se considera la mejor opción, pedir los permisos correspondientes a los propietarios de los derechos de autor, incluso cuando el test solo se vaya a emplear con fines de investigación (ver las directrices de la ITC sobre el uso de los test en investigación; ITC, 2014 – traducidas al español en https://www.cop.es/pdf/ITC2015-Investigacion.pdf - Muñiz, Hernández y Ponsoda, 2015).

Aceptable=Excelente: Se obtiene permiso escrito de los propietarios de los derechos del test.

* Es necesario tener en cuenta que, en muchos casos, el propietario de los derechos del test no es el autor del test sino la editorial o casa distribuidora. Además, a la hora de obtener los permisos, es necesario tener en cuenta las siguientes cuestiones: si los editores/propietarios del copyright permiten cambios en la estructura del test, basados, por ejemplo, en los resultados de un estudio piloto; si aceptarían cambios en el contenido del test y no solo en la traducción; si están de acuerdo en que los estudios sobre el test y sus posibles modificaciones puedan publicarse en otro país; y si aceptan que los baremos, si los hubiera, se presenten de un modo distinto al original, en función de los resultados derivados del proceso de adaptación del test.

En ciertos proyectos internacionales (e.g., PISA, TIMMS), es frecuente que los test se construyan simultáneamente para ser usados en distintas lenguas y culturas. En caso del desarrollo simultáneo del test, este criterio no sería aplicable, ya que no hay un test original que deba ser traducido y adaptado para su uso en otra población.

DP3-1 (G3, C4): Si la adaptación se considera la mejor opción, evaluar las posibles diferencias culturales y lingüísticas antes de comenzar el proceso de adaptación. Estas diferencias deben ser consideradas en la versión adaptada, con el fin de prevenir sesgos y diseñar estudios que permitan controlar los potenciales sesgos (Arnold y Smith, 2013; Pena, 2007).

Excelente: Las diferencias culturales y lingüísticas (uso de distintos formatos de ítems, familiaridad con los materiales y el lenguaje, conceptos emic y etic, estilos de vida, etc.) detectadas según la opinión de los expertos (tanto en la lengua como en la cultura) y las evidencias empíricas (por ejemplo, entrevistas, observaciones o encuestas a individuos de la población diana) son sistematizadas y documentadas.

Aceptable: Las diferencias culturales y lingüísticas más importantes detectadas según la opinión de los expertos (tanto en la lengua como en la cultura) se resumen y documentan. Sin embargo, no se dispone todavía de evidencia empírica.

2. DIRECTRICES DE DESARROLLO

DD1 "Asegurarse, mediante la selección de expertos cualificados, de que el proceso de [traducción y] adaptación tiene en cuenta las diferencias lingüísticas, psicológicas y culturales entre las poblaciones de interés"

DD2 "Utilizar diseños y procedimientos racionales apropiados para asegurar la adecuación de la adaptación del test a la población a la que va dirigido"

DD3. Ofrecer "evidencias que garanticen que las instrucciones del test y el contenido de los ítems tienen un significado similar en todas las poblaciones a las que va dirigido el test"

DD4 Ofrecer "evidencias que garanticen que el formato de los ítems, las escalas de respuesta, las reglas de corrección, las convenciones utilizadas, las formas de aplicación y demás aspectos son adecuados para todas las poblaciones de interés"

DD5. "Recoger datos mediante estudios piloto sobre el test adaptado, y efectuar análisis de ítems y estudios de fiabilidad y validación que sirvan de base para llevar a cabo las revisiones necesarias y adoptar decisiones sobre la validez del test adaptado"

NOTA: Antes de comenzar el proceso de adaptación del test, los investigadores deben asegurar que cumplen con los principios éticos que rigen la investigación con humanos y cerciorarse de que los participantes implicados en el proceso de adaptación dan su consentimiento informado siguiendo los modelos establecidos en cada país.

Operacionalización

DD1-1 (G4, C5): Formar un equipo multidisciplinar compuesto por: a) traductores profesionales para traducir el test de la lengua original a la lengua de adaptación (cuando sea necesaria la traducción) y que tengan cierto conocimiento de las culturas implicadas, b) expertos en el constructo a medir, c) expertos en las culturas implicadas, y d) expertos en la construcción de test. En algunos casos, un mismo miembro del equipo puede ser experto en más de uno de estos aspectos; por ejemplo, en las lenguas y culturas, en el constructo y las culturas, etc. (Epstein, Santo, y Guillemin, 2015).

*Excelente**: El equipo incluye un mínimo de cinco expertos: dos traductores profesionales con conocimiento de las culturas implicadas, que han recibido instrucción sobre la construcción de ítems, un experto en el constructo a medir, un experto en las culturas implicadas y un experto en construcción de test.

Aceptable El equipo incluye un mínimo de tres expertos: dos traductores profesionales con conocimiento de las culturas implicadas y un experto en el constructo a medir y/o en la construcción de test. En este caso se permite un cierto solapamiento entre categorías: los traductores pueden tener experiencia en la construcción de test, o tener conocimientos sobre el constructo a medir, por ejemplo.

*Es recomendable que haya más expertos cuanto más importantes sean las decisiones que se vayan a tomar a partir de las puntuaciones de los test, cuando estas requieran la comparación de individuos o grupos de distintas culturas. Cuando la adaptación no requiera traducción, no será necesario incluir traductores profesionales en el equipo (por ejemplo, cuando un test desarrollado en España vaya a ser adaptado para su administración en México). Sin embargo, la equivalencia lingüística sigue siendo crucial, por lo que es necesario seguir

contando con un equipo multidisciplinar. Para asegurar la calidad del equipo, el procedimiento para seleccionar los expertos, así como los títulos y experiencia requeridos, deben quedar documentados. Finalmente, si la adaptación del test requiere ser aplicada a poblaciones con necesidades especiales, el equipo debe incorporar profesionales adicionales (e.g., psicólogos o profesionales de educación especial -ver Leong, Bartram, Cheung, Geisinger e Iliescu, 2016).

DD2-1 (G5, C6): Usar alguno de los diseños de traducción recomendados y justificar la elección. La traducción hacia adelante, hacia atrás (o retro-traducción) o la traducción simultánea son posibles alternativas, dependiendo del propósito de la traducción, del alcance del proyecto, del número de culturas implicadas y de si es o no necesario comparar las puntuaciones de personas que pertenecen a las distintas culturas implicadas. Para los diseños de traducción hacia adelante o hacia atrás, se requiere que se empleen al menos dos traductores (o equipos de traductores) independientes (Hambleton, 2005). Si un test se construye desde sus inicios con el fin de ser aplicado transculturalmente, es posible el desarrollo simultáneo/concurrente de las múltiples versiones del test desde el principio.

Excelente: Se realiza una traducción hacia adelante mediante distintos traductores independientes (Hagell, Hedin, Meads, Nyberg, y McKenna, 2010) o se combinan varios diseños de traducción para obtener la versión inicial del test adaptado (Wild et al., 2005). La razón para preferir la traducción hacia adelante (de la lengua original a la lengua diana) a la retro-traducción es que las posibles discrepancias existentes son identificadas y revisadas directamente en la lengua diana o versión adaptada del test (ITC, 2017). En estudios de evaluación cross-cultural de gran alcance (como PISA) se pueden usar diferentes versiones del test en distintas lenguas, que serán traducidas de forma independiente para, tras la revisión y reconciliación de las distintas propuestas, llegar a la versión diana (Grisay, 2003). Además de permitir la identificación y revisión de las posibles discrepancias directamente en la versión adaptada que emplea la lengua diana, la práctica de usar más de una lengua de origen para la traducción ayuda a minimizar el impacto de la cultura ligada a la lengua original (ITC, 2017).

Aceptable: Se realiza una traducción hacia atrás siguiendo las recomendaciones de este diseño (Brislin, 1986).

DD2-2 (G5, C7): Contar con varios traductores que trabajan de forma independiente y constituir un comité de expertos que revisen y comparen las traducciones propuestas, con el fin de recoger sus opiniones sobre las versiones, resolver las discrepancias existentes, y proponer una versión consensuada (Epstein, Osborne, Elsworth, Beaton y Guillemin, 2015; Koller et al., 2012).

Excelente: Al menos dos traductores trabajan en cada fase del diseño, y un comité independiente de traductores cualificados y expertos (en la cultura, el constructo, y en medición mediante test) revisa las distintas versiones y trabaja con los traductores originales para resolver las posibles discrepancias.

Aceptable: Al menos dos traductores trabajan en cada fase del diseño, y un traductor independiente con conocimiento en las culturas implicadas, revisa las distintas versiones y trabaja con los traductores originales para resolver las posibles discrepancias entre las versiones.

DD3-1 (G6, C8): Asegurar que las instrucciones del test son claras y comprensibles, empleando un lenguaje familiar para la población diana.

Excelente: El equipo de expertos ha revisado, aprobado y documentado la adecuación de las instrucciones. Además, se ha realizado algún estudio piloto o pre-test (e.g., entrevistas cognitivas; ver Levin et al., 2009; Padilla y Benítez, 2014), cuyos resultados apoyan la adecuación de las instrucciones.

Aceptable: El equipo de expertos ha revisado, aprobado y documentado la adecuación de las instrucciones.

DD3-2 (G6, C9): Asegurar que el contenido de los ítems resulta claro y se expresa con los mismos niveles de familiaridad y dificultad en la lengua original y la lengua diana. Los elementos lingüísticos que pudieran dificultar la comprensión de la versión traducida, como palabras con diferentes significados, dobles

negaciones, etc. deben evitarse. Los elementos no verbales (imágenes o dibujos) deben estar contextualizados teniendo en cuenta la cultura de la población (Hambleton y Zenisky, 2011; van der Vijver y Tanzer, 2004).

*Excelente**: El equipo de expertos ha revisado, aprobado y justificado la adecuación y comparabilidad de los enunciados de los ítems. Además, se ha realizado algún estudio piloto o pre-test (e.g., entrevistas cognitivas) con muestras de las poblaciones de interés (idealmente bilingües), cuyos resultados indican que no hay problemas de comprensión con los enunciados propuestos y que éstos son interpretados de forma similar a los ítems originales.

*Aceptable**: El equipo de expertos ha revisado, aprobado y justificado la adecuación y comparabilidad de los enunciados de los ítems.

* Se recomienda que los expertos usen listados de verificación de la calidad de los ítems adaptados, como el listado que proponen Hambleton y Zenisky (2011), considerando distintos aspectos relevantes en función del tipo de test a adaptar (este listado también está disponible en español – Muñiz et al., 2013)

DD4-1 (G7, C10): Intentar que el formato de los ítems, las opciones de respuesta, las rúbricas de puntuación si las hubiera, y el modo de aplicación del test sean similares en las distintas versiones (Beaton, Bombardier, Guillemin, y Ferraz, 2000).

Excelente: Los expertos están de acuerdo (y justifican razonadamente) que el formato de los ítems, las opciones de respuesta, las rúbricas de puntuación si las hubiera, y el modo de aplicación son muy similares en las distintas versiones, sin que se observen diferencias importantes.

Aceptable: Los expertos observan algunas diferencias significativas en el formato de los ítems, las opciones de respuesta, etc., pero están de acuerdo (y justifican razonadamente) que las diferencias no alteran substancialmente el significado o la dificultad de los ítems

DD4-2 (G7, C11): Asegurar que la población diana está suficientemente familiarizada con los procedimientos empleados (formato de los ítems, escalas de respuesta, rúbricas de puntuación (si las hubiera), y modo de aplicación) en el test adaptado (Hambleton, Merenda, y Spielberg, 2005; van der Vijver y Leung, 2011).

Excelente: Hay estudios previos en los que se han empleado los procedimientos implicados (formato de ítems, modo de aplicación, etc.) y no hay indicios de que exista ningún problema al respecto. Además, se cuenta con un estudio piloto que muestra que, o bien los individuos pertenecientes a la población diana no tienen problemas en el uso de los procedimientos implicados, o bien, si se detectaran dificultades, estas desaparecen tras presentar un número suficiente de ítems de prueba, que permitan a los participantes familiarizarse con todos los aspectos de los procedimientos implicados

Aceptable: O bien hay estudios previos en el que se han empleado los procedimientos implicados, o bien se realiza un estudio piloto en el que no hay indicios de que exista ningún problema al respecto. Si en el estudio piloto se detectan dificultades se presenta un número suficiente de ítems de prueba, para asegurar que los participantes se familiarizan adecuadamente con todos los aspectos de los procedimientos implicados

DD5-1 (G8, C12): Comprobar la calidad psicométrica (análisis de ítems, fiabilidad y validez) de las puntuaciones del test adaptado en una muestra piloto de la población diana y, considerando los resultados, hacer las revisiones necesarias para mejorar la versión final del test. Las muestras del estudio piloto deben ser suficientemente grandes para realizar los análisis estadísticos correspondientes.

*Excelente**: Se ha realizado un análisis de los ítems (dificultad, discriminación, y, cuando sea necesario, análisis de distractores) y se ha estimado la fiabilidad de las escalas (o la función de información, si se aplica la Teoría de la Respuesta a los Ítems (TRI)). Además, se ha realizado al menos un estudio de

validación. Este puede centrarse en las relaciones entre las puntuaciones del test y otras variables relevantes (e.g., inter-correlaciones entre distintas dimensiones del test, o correlaciones con variables de la red nomológica, para evaluar la validez referida a un criterio, validez convergente, discriminante, o diferencias grupales). También puede centrarse en las evidencias de validez basadas en la estructura interna del test**. Los resultados de los análisis alcanzan niveles de excelencia de acuerdo con criterios estándar (ver, por ejemplo, los modelos de revisión de test de la EFPA (*European Federation of Psychological Associations)* y del COP (Colegio Oficial de Psicólogos) – ver Evers et al., 2013; Hernández, Ponsoda, Muñiz, Prieto y Elosua, 2016, o Furr, 2017 y Mellenbergh, 2011). Si hay ítems problemáticos se eliminan o mejoran para alcanzar resultados óptimos.

*Aceptable**: Se ha realizado un análisis de los ítems (dificultad, discriminación, y, cuando sea necesario, análisis de distractores) y se ha estimado la fiabilidad de las escalas (o la función de información, si se aplica TRI). Los resultados de los análisis alcanzan niveles de excelencia de acuerdo con criterios estándar (como por ejemplo del modelo de la EFPA y del COP). Si hay ítems problemáticos se eliminan o mejoran para alcanzar resultados aceptables.

* Las evidencias de validez basadas en el contenido de los ítems y/o las basadas en el proceso de respuesta (e.g., entrevistas cognitivas) se han considerado en DD3-2 (G6, C9).

** Cuando existen estudios sobre la equivalencia métrica de las puntuaciones a través de distintos grupos o culturas, se obtienen evidencias de validez basadas en la estructura interna, al evaluar si la estructura es la esperada y es invariante a través de los grupos de comparación.

3. DIRECTRICES DE CONFIRMACIÓN [ANALISIS EMPÍRICO]

C1 "Definir las características de la muestra que sean pertinentes para el uso del test, y seleccionar un tamaño de muestra suficiente que sea adecuado para las exigencias de los análisis empíricos"

C2 "Ofrecer información empírica pertinente sobre la equivalencia del constructo, equivalencia del método y equivalencia entre los ítems en todas las poblaciones implicadas"
*C3 "Recoger información y evidencias [*que apoyen el uso de los baremos], *la fiabilidad y la validez de la versión adaptada del test en las poblaciones implicadas".*
C4 Cuando se comparen puntaciones entre distintas versiones del test, emplear diseños de equiparación y análisis de datos adecuados que garanticen la comparabilidad.

Operacionalización

C1-1 (G9, C13): Asegurar que la muestra diana sea lo suficientemente grande como para llevar a cabo los análisis estadísticos necesarios y para representar adecuadamente a la población*.

Excelente: La muestra diana se ha elegido utilizando un muestreo probabilístico y su tamaño es apropiado según el tamaño de la población y la variabilidad de las características de interés. Además, la muestra es lo suficientemente grande como para llevar a cabo los análisis estadísticos y obtener estimaciones estables de los parámetros investigados en el estudio.

Aceptable: El tamaño de la muestra objetivo es apropiado de acuerdo con el tamaño y la variabilidad de las características medidas en la población. La muestra no se ha elegido utilizando el muestreo probabilístico, pero se han considerado las características sociodemográficas más importantes de la población. Además, la muestra es lo suficientemente grande como para llevar a cabo los análisis estadísticos y obtener estimaciones estables de los parámetros investigados en el estudio.

* Es importante tener en cuenta que algunas técnicas, como la TRI o el análisis factorial confirmatorio (AFC), pueden requerir muestras de gran tamaño (Byrne y Van de Vijver, 2014; DeMars, 2010; Gagne y Hancock, 2006; Wolf, Harrington, Clark, y Miller, 2013).

C1-2 (G9, C14): Cuando el interés se centra en las comparaciones transculturales, asegurar que la muestra original y la muestra diana son comparables para todas las

variables pertinentes, excepto el idioma y/o contexto cultural (van der Vijver y Leung, 1997).

Excelente: La muestra diana tiene las mismas características que la muestra original (es decir, no hay diferencias significativas en variables relevantes: sociodemográficas, académicas, clínicas, etc.), excepto el idioma y/o contexto cultural.

Aceptable: Las diferencias existentes entre la muestra original y la muestra diana en variables relevantes, se han controlado usando diseños de emparejamiento o procedimientos estadísticos adecuados.

C2-1 (G10, C15): Cuando haya interés en comparar la población original con la población diana, utilizar previamente procedimientos estadísticos para asegurar la equivalencia del constructo en las distintas poblaciones (van der Vijver y Poortinga, 2004; van der Vijver y Tanzer, 2004).

Excelente: Se han aplicado procedimientos basados en la estructura interna del test (por ejemplo, análisis factorial) y en la red nomológica (mediante la comparación estadística del patrón de relaciones entre el constructo y otras variables relevantes) con tamaños de muestra adecuados. Los resultados obtenidos en todos los procedimientos apoyan la equivalencia del constructo.

Aceptable: Se ha aplicado un procedimiento adecuado, basado en la estructura interna del test o en el análisis de la red nomológica, con tamaños de muestra adecuados y los resultados apoyan la equivalencia del constructo.

C2-2 (G10, C16): Cuando haya interés en comparar la población original con la población diana, comprobar la equivalencia de método (características del instrumento, proceso de administración y características de la muestra).

Excelente: Se ha logrado la excelencia en los criterios relativos al sesgo de la muestra (C1), las características del instrumento (TD4) y el proceso de administración (A1 y A2)*. Se realizan análisis adicionales para comprobar si

las potenciales amenazas metodológicas identificadas en las fases de diseño, desarrollo y administración están afectando a los resultados finales. Los resultados muestran que los efectos del método no son un problema.

Aceptable: Se alcanza un nivel aceptable (como mínimo) para los criterios relativos al sesgo de la muestra (C1), las características del instrumento (TD4) y el proceso de administración (A1 y A2)*. Se realizan análisis adicionales para comprobar si las potenciales amenazas metodológicas identificadas en las fases de diseño, desarrollo y administración afectan los resultados finales. Si los resultados muestran que los efectos del método son un problema, se controla su influencia.

* El nivel de cumplimiento de este criterio puede evaluarse completamente después de evaluar los criterios A1 y A2 de las directrices de administración.

C2-3 (G10, C17): Cuando haya interés en comparar las poblaciones objetivo y de referencia, evaluar el Funcionamiento Diferencial de los Ítems (DIF -*Differential Item Functioning*) entre los grupos culturales que se van a comparar utilizando procedimientos estadísticos adecuados al formato del ítem, el tamaño de la muestra y la dimensionalidad del test* (Hidalgo y Gómez-Benito, 2010; Sireci y Rios, 2013).

Excelente: Se aplica la técnica estadística más adecuada para la detección de DIF (considerando el formato del ítem, el tamaño de la muestra, etc.) utilizando un procedimiento de purificación del criterio de equiparación. Además de utilizar pruebas de significación, se reportan indicadores del tamaño del efecto para el DIF e indicadores del Funcionamiento Diferencial del Test (FDT).

Aceptable: Se utiliza la técnica estadística más adecuada para la detección de DIF (considerando el formato del ítem, el tamaño de la muestra, etc.). Además de utilizar pruebas de significación, se reportan indicadores del tamaño del efecto para el DIF.

* Los resultados de estos análisis deben tenerse en cuenta en las etapas posteriores del proceso de adaptación (véase, en particular, C4-1 y la sección sobre escalas de puntuación y directrices de interpretación).

C2-4 (G10, C18): En caso de que el DIF se detecte en niveles significativos, llevar a cabo análisis para comprender las causas del DIF (por ejemplo, efectos lingüísticos o de método) entre las culturas (Benitez, Padilla, Hidalgo, y Sireci, 2016; Gómez-Benito, Sireci, Padilla, Hidalgo, y Benitez, 2018).

Excelente: Se utilizan métodos mixtos que combinan enfoques cualitativos (por ejemplo, entrevistas cognitivas o jueces expertos) y cuantitativos (por ejemplo, experimentos o cuasiexperimentos), para comprender las causas del DIF.

Aceptable: Se llevan a cabo estudios cualitativos o cuantitativos para comprender las causas del DIF.

C3-1 (G11, C19): Asegurar que el tipo de indicadores de fiabilidad reportados son los adecuados para el tipo de test, utilizando análisis estadísticos y tamaños muestrales adecuados. Los valores obtenidos deben ser satisfactorios y acompañarse del error típico de medición (AERA, APA, y NCME, 2014; Evers et al., 2013; Terwee et al., 2012).

Excelente: Dependiendo del propósito del test, se proporciona evidencia sobre la consistencia interna, la estabilidad de la medida y/o el acuerdo entre calificadores (para más información, véase las normas APA, 2014). También se proporciona información sobre el error típico de medida, obtenido por medio de la teoría clásica del test o la TRI. Los resultados muestran valores buenos o excelentes cuando se considera la importancia de las decisiones que deben tomarse a partir del test (véase el modelo de revisión de test de la EFPA; Evers et al., 2013).

Aceptable: Dependiendo del propósito del test, se proporciona evidencia sobre la consistencia interna, la estabilidad de la medida y/o el acuerdo entre calificadores (para más información, véase las normas APA, 2014). También

se proporciona información sobre el error típico de medida, obtenido por medio de la teoría clásica del test o la teoría de respuesta al ítem. Los resultados muestran valores aceptables cuando se considera la importancia de las decisiones que deben tomarse a partir del test (véase el modelo de revisión de test de la EFPA; Evers et al., 2013).

C3-2 (G11, C20): Proporcionar evidencias de validez coherentes con el uso previsto de los resultados del test, utilizando análisis estadísticos y tamaños muestrales adecuados (AERA, APA, y NCME, 2014; Evers et al., 2013; Terwee et al., 2012).

Excelente: Se formulan hipótesis/objetivos de validación para diferentes tipos de evidencias de validez. Se informa de las evidencias basadas en la validez del contenido (parcialmente comprobadas mediante el criterio "excelente" para las directrices PC2 y TD1 a TD3), la estructura interna del test y las relaciones con otras variables. Cuando sea pertinente, también se considerará la evidencia basada en el proceso de respuesta (parcialmente comprobada a través del criterio de "aceptable" para la directriz TD3) y/o las consecuencias del uso del test. Los resultados obtenidos considerando todas las fuentes de evidencia apoyan el uso previsto del test.

Aceptable: Se formulan hipótesis/objetivos de validación para el uso previsto de los resultados del test. Los resultados de la evidencia recogida apoyan el uso previsto del test.

C3-3 (G11, C21): Asegurar y verificar que las normas son adecuadas para interpretar los resultados del test en la población diana.

Excelente: Se han elaborado normas específicas para la población diana. La muestra se ha elegido utilizando métodos de muestreo probabilísticos, es lo suficientemente grande como para garantizar la representatividad y se reporta el error de muestreo. Además, cuando sea necesario, aparte de proporcionar las normas de interpretación generales, se elaboran normas específicas para tener

en cuenta las variables relevantes (edad, sexo, etc.) o poblaciones específicas (por ejemplo, clínicas o con necesidades especiales).

Aceptable: O bien hay evidencia de que las normas elaboradas en la versión original pueden utilizarse para la población diana, o bien se han elaborado normas específicas para la población diana. En este último caso, la muestra no se ha elegido siguiendo el método de muestreo probabilístico, pero se ha seleccionado considerando las características sociodemográficas más importantes de la población.

C4-1 (G12, C22): Cuando el objetivo es la evaluación intercultural o interlingüística y algunos ítems funcionan diferencialmente, se utilizan diseños de equiparación y procedimientos de análisis apropiados antes de la comparación (Angoff, 1984; Dorans, Pommerich, y Holland, 2007; Kolen y Brennan, 2004).

Excelente: Para elegir los ítems de anclaje que servirán para equiparar las muestras, se combinan criterios de juicio (es decir, ítems que los expertos consideran más fáciles para traducir y adaptar) y criterios empíricos (es decir, utilizando muestras bilingües o un subconjunto de ítems que han demostrado estar libres de DIF en el análisis de DIF). Los análisis de datos de equiparación y los tamaños muestrales son adecuados.

Aceptable: Para elegir los ítems de anclaje que servirán para equiparar las muestras, se utilizan criterios de juicio (es decir, ítems que los expertos consideran más fáciles para traducir y adaptar) o criterios empíricos (es decir, utilizando muestras bilingües o un subconjunto de ítems que han demostrado estar libres de DIF en el análisis de DIF). Los análisis de datos de equiparación y los tamaños muestrales son adecuados.

4. DIRECTRICES SOBRE LA APLICACIÓN

A1 "Preparar los materiales y las instrucciones para la aplicación de modo que minimicen cualquier diferencia cultural y lingüística que pueda ser debida a los procedimientos de aplicación y a los formatos de respuesta, y que puedan afectar a la validez de las inferencias derivadas de las puntuaciones".

A2 "Especificar las condiciones de aplicación del test que deben seguirse en todas las poblaciones a las que va dirigido"

Operacionalización

A1-1 (G13, C23): Para todos los materiales e instrucciones de aplicación del test, comprobar el cumplimiento de los requisitos establecidos en las directrices de desarrollo (TD3 a TD5). Para prevenir posibles problemas en la administración del test adaptado en la población diana, se ha de tener en cuenta la experiencia acumulada en la aplicación de la versión original del test en la población de origen.

Excelente: Los criterios establecidos en las directrices de desarrollo (TD3 a TD5) se cumplen al nivel de "Excelente".

Aceptable: Los criterios establecidos en las directrices de desarrollo (TD3 a TD5) se cumplen como mínimo al nivel de "Aceptable".

A2-1 (G14, C24): Cuando se vayan a realizar comparaciones transculturales, asegurar que las condiciones de administración del test (modo de administración, limitación de tiempo, información sobre el propósito de la prueba, etc.) están estandarizadas a través de los grupos. En caso en que fuera necesario introducir cambios, evaluar el posible impacto de las diferentes condiciones de administración del test.

Excelente: Las condiciones de administración del test son las mismas en las diferentes culturas, o hay un estudio piloto que muestra que los cambios introducidos no afectan a los resultados.

Aceptable: Las diferencias en las condiciones de administración son inevitables pero están descritas claramente, lo que permite hacer comparaciones entre las poblaciones de manera prudente. Se aportan algunas evidencias que indican que estas diferencias no tienen un impacto considerable.

A2-2 (G14, C25): Asegurar que los entrevistadores o los administradores de los test tengan las credenciales necesarias según el tipo de test que se vaya a aplicar. Los administradores del test deben presentar una declaración firmada por la que se comprometen a llevar a cabo sus actividades de conformidad con el código ético y los principios de la práctica profesional establecidos por las asociaciones y órganos profesionales nacionales competentes.

Excelente: Todos los administradores tienen las credenciales requeridas (si las hubiera) y experiencia documentada en la aplicación del tipo de test que se ha adaptado. Los administradores de la prueba han firmado el compromiso, mencionado anteriormente, de llevar a cabo sus actividades de acuerdo con los códigos de conducta éticos y profesionales.

Aceptable: Los administradores de la prueba tienen las credenciales requeridas (si las hubiera). Es posible que todos o algunos de ellos no tengan experiencia en la administración del tipo de test adaptado, pero se les proporciona formación y se evalúa su resultado. Sólo se seleccionan como administradores de la prueba a las personas cuyo desempeño tras la formación es adecuado. Los administradores de la prueba han firmado el compromiso, mencionado anteriormente, de llevar a cabo sus actividades de acuerdo con códigos de conducta éticos y profesionales.

5. DIRECTERICES SOBRE PUNTUACIÓN E INTERPRETACIÓN

SSI1 "Interpretar las diferencias de las puntuaciones entre los grupos teniendo en cuenta la información demográfica pertinente."
SSI2 "Comparar las puntuaciones entre poblaciones únicamente en el nivel de invarianza establecida para la escala de puntuación utilizada en las comparaciones."

Operacionalización

Antes de valorar la directriz SSI1, sobre las diferencias de las puntuaciones entre los grupos, es necesario determinar si las puntuaciones son comparables de acuerdo con la directriz SSI2. Por lo tanto, comenzamos con SSI2 y luego procedemos con SSI1.

SSI2-1 (G16, C26): Antes de comparar las puntuaciones individuales de personas pertenecientes a diferentes culturas, y/o las puntuaciones promedio de grupos de diferentes culturas, se debe asegurar que la equivalencia métrica (i.e. la ausencia de DIF) ha sido evaluada y se cumple, al menos para un número relevante de ítems* (Byrne, 2008; Byrne y van der Vijver, 2017; Dimitrov, 2010; van der Vijver y Leung, 2011).

Excelente: La invarianza de la medida ha sido evaluada y confirmada: las diferencias en los parámetros no son estadísticamente significativas, o son triviales, según los tamaños del efecto.

Aceptable: La invarianza de la medida ha sido evaluada y se cumple únicamente la invarianza parcial. En este caso, sólo los ítems invariantes deben incluirse en las comparaciones de grupo, siempre que la validez de contenido no se vea comprometida. Alternativamente deben utilizarse puntuaciones corregidas (tras aplicar un diseño de equiparación adecuado).

* Téngase en cuenta que, si el foco de interés es únicamente la comparabilidad de las relaciones entre el constructo de interés y otros constructos relevantes, sólo es necesario aportar evidencias de la invariancia métrica (i.e. de las relaciones entre el ítem y el constructo a medir) a través de los grupos.

SSI1-1 (G15, C27): Cuando las comparaciones entre puntuaciones estén justificadas sobre la base del análisis de invarianza de la medida, interpretar las diferencias interculturales observadas (si las hubiera), según la información sistematizada y documentada en PC3-1 (G3, C4) en relación con la distancia cultural y lingüística. Para comprender las diferencias en las puntuaciones observadas, debe considerarse el papel de estas variables (por ejemplo, religiosidad, individualismo, diferentes tendencias de respuesta) (van der Vijver y Leung, 2011).

Excelente: Se ha considerado la información sistematizada y documentada en PC3-1 (G3, C4) y se ha analizado empíricamente el papel de las variables clave.

Aceptable: La información resumida en PC3-1 (G3, C4) se ha empleado para proporcionar explicaciones razonables de las diferencias obtenidas.

6. DIRECTRICES SOBRE LA DOCUMENTACIÓN

Doc1 "Proporcionar documentación técnica que recoja cualquier cambio en el test adaptado, incluyendo la información y las evidencias sobre la equivalencia entre las versiones adaptadas."

Doc2 "Proporcionar documentación a los usuarios con el fin de garantizar un uso correcto del test adaptado en la población a la que va dirigido."

Doc1-1 (G17, C28): Crear una serie de documentos y ponerlos a disposición de las personas y sectores interesados, proporcionando información sobre las siguientes cuestiones:

a) Evidencias (teóricas y empíricas) de que el constructo de interés es relevante en la población destino y tiene el mismo significado que en la población de origen (tal y como se ha considerado en PC2-1 (G2, C1) y PC2-2 (G2, C2)).
b) Distancia cultural y diferencias lingüísticas entre las culturas basadas en la opinión de los expertos (en el idioma y cultura de la población diana) y en evidencias empíricas (entrevistas, observación o encuestas a personas de la población diana, etc.) (tal como se considera en PC3-1 (G3, C4)).
c) Las medidas adoptadas para prevenir el sesgo (por ejemplo, descentralización cultural, diseño de traducción, elaboración de un protocolo de aplicación detallado y estandarizado entre las culturas, etc.) (véase A1 y A2).
d) Las desviaciones respecto del test original y de las condiciones originales de aplicación (p.e., número de ítems, modo de administración, duración), indicando las razones de los cambios y, cuando el propósito principal sea comparar las puntuaciones entre culturas, las evidencias recogidas de que esos cambios no comprometen gravemente la comparabilidad entre las puntuaciones (véase A1 y A2).
e) Información sobre los miembros del equipo que han trabajado en el proceso de adaptación y sobre el diseño de adaptación, justificando el diseño concreto que se ha seleccionado.

f) Cuando se requiera comparar las puntuaciones de personas o grupos de diferentes culturas, información sobre las características de la población original y de destino, el procedimiento de selección de las muestras, las características de las muestras, sus principales diferencias (si las hubiera) y la posible influencia de dichas diferencias sobre la comparabilidad de las puntuaciones entre las culturas.
g) Información técnica sobre las propiedades psicométricas de las puntuaciones obtenidas con el test adaptado (análisis de ítems, fiabilidad, y evidencias de validez).
h) Cuando el objetivo de la adaptación sea comparar individuos de diferentes culturas, documentación técnica que incluya los resultados de los estudios de equivalencia métrica realizados y las conclusiones sobre el nivel al que pueden compararse las puntuaciones obtenidas en las distintas culturas.

Excelente: La documentación incluye información exhaustiva sobre todas las cuestiones mencionadas. Esta información es necesaria para que los usuarios determinen la idoneidad de los cambios que se han hecho y para que los investigadores puedan reproducir el proceso de adaptación. Se indica dónde está disponible la documentación y cómo se puede acceder a ella.

Aceptable: La documentación incluye información suficiente sobre las cuestiones mencionadas para juzgar la idoneidad del proceso de adaptación. Se indica dónde está disponible la documentación y cómo se puede acceder a ella.

Doc2-1 (G18, C29): Asegurar que los materiales y la documentación que acompañan al test (p.e., el manual del test) sean claros (instrucciones, descripción del ámbito de aplicación, ejemplos prácticos de su uso, etc.) para garantizar que el test sea adecuado para la población a la que se dirige, que la administración del test esté estandarizada y que las puntuaciones se puedan interpretar adecuadamente (véanse las secciones de aplicación y puntuación).

Excelente: Se han tenido en cuenta las Directrices de la ITC sobre el uso de los test, así como los estándares aceptados para evaluar la calidad de los test (de acuerdo, por ejemplo, con el modelo de revisión de test de la EFPA o los

Standards for Educational and Psychological Testing (AERA, APA y CCME, 2014), entre otros). Los resultados muestran valores buenos o excelentes (Evers et al., 2013)

Aceptable: Se han tenido en cuenta las Directrices de la ITC sobre el uso de los test, así como los estándares aceptados para evaluar la calidad de los test (de acuerdo, por ejemplo, con el modelo de revisión de test de la EFPA o los *Standards for Educational and Psychological Testing* (AERA, APA y CCME, 2014), entre otros). Los resultados muestran valores adecuados (Evers et al., 2013).

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, y CCME) (2014). *Standards for educational and psychological testing*. Washington, DC, USA: American Psychological Association.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service.

Arnold, B. R., y Smith, J. L. (2013). Methodologies for test translation and cultural equivalence. En F. Paniagua y A.M. Yamada (Eds.), *Handbook of multicultural mental health: Assessment and treatment of diverse populations* (2nd ed.). San Diego, CA, US: Elsevier Academic Press.

Beaton, D. E., Bombardier, C., Guillemin, F., y Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, *25*(24), 3186-3191.

Benítez, I., Padilla, J. L., Hidalgo Montesinos, M. D., y Sireci, S. G. (2016). Using mixed methods to interpret differential item functioning. *Applied Measurement in Education*, *29*(1), 1-16.

Brislin, R.W. (1986). The wording and translation of research instruments. En W.J. Lonner y J.W. Berry (Eds.), *Field methods in cross-cultural psychology* (pp. 137-164). Newbury Park, CA: Sage Publications

Byrne, B. M. (2008). Testing for multigroup equivalence of measuring instrument: A walk through the process. *Psicothema, 20,* 872-882.

Byrne, B. M., y van der Vijver, F. J. R. (2014). Factorial structure of the Family Values Scale from a multilevel-multicultural perspective. *International Journal of Testing, 14,* 168-192.

Byrne, B. M., y van der Vijver, F. J. R. (2017). The maximum likelihood alignment approach to testing for approximate measurement invariance: A paradigmatic cross-cultural application. *Psicothema, 29,* 539-551.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counselling and Development, 43,* 121-149.

Dorans, N. J., Pommerich, M., y Holland, P.W. (Eds.) (2007). *Linking and aligning scores and scales*. New York, NY: Springer.

Epstein, J., Osborne, R. H., Elsworth, G. R., Beaton, D. E., y Guillemin, F. (2015). Cross-cultural adaptation of the Health Education Impact Questionnaire: Experimental study showed expert committee, not back-translation added value. *Journal of Clinical Epidemiology, 68,* 360-369.

Epstein, J., Santo, R.M., y Guillemin, F. (2015). A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. *Journal of Clinical Epidemiology, 68,* 435-441.

Evers, A., Muñiz. J., Hagemeister, C., Hstmælingen, A., Lindley, P., Sjöberg, A., y Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema, 25* (3), 283-291. doi: 10.7334/psicothema2013.97

Furr, R. M. (2017). *Psychometrics: An introduction*. Sage Publications.

Gagne, P., y Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research, 41,* 65-83.

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., y Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema*, *30*(1), 104-109.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20* (2), 225-240.

Hagell, P., Hedin, P. J., Meads, D. M., Nyberg, L., y McKenna, S. P. (2010). Effects of method of translation of patient-reported health outcome questionnaires: A

randomized study of the translation of the Rheumatoid Arthritis Quality of Life (RAQoL) instrument for Sweden. *Value in Health*, *13*(4), 424-430.

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G. y Elosua, P. (2016). Revisión del modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 37,* 192-197.

Hambleton, R. K. (2005). Issues, designs and technical guidelines for adapting tests into multiple languages and cultures. En R. K. Hambleton, P. F. Merenda, y S. D. Spielberger (eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). New Jersey: Lawrence Erlbaum Associates.

Hambleton, R. K., Merenda, P. F., y Spielberger, S. D. (eds.) (2005), *Adapting educational and psychological tests for cross-cultural assessment*. New Jersey: Lawrence Erlbaum Associates.

Hidalgo, M.D., y Gómez-Benito, J. (2010). Education measurement: Differential item functioning. In B. McGaw, P. Peterson, y E. Baker (Eds.), *International Encyclopedia of Education*, 3rd edition, vol. 4, pp 36-44. Elsevier Science y Technology.

International Test Commission. (2014). ITC statement on the use of tests and other assessment instruments for research purposes. [www.intestcom.org]

International Test Commission. (2017). The ITC Guidelines for Translating and Adapting Tests (Second edition). [www.intestcom.org]

Kolen, M. J., y Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Koller, M., Kantzer, V., Mear, I., Zarzar, K., Martin, M., Greimel, E., ... y ISOQOL TCA-SIG. (2012). The process of reconciliation: Evaluation of guidelines for translating quality-of-life questionnaires. *Expert Review of Pharmacoeconomics y Outcomes Research*, *12*(2), 189-197.

Leong, F.T.L., Bartram, D., Cheung, F.M., Geisinger, K.F, y Iliescu, C. (2016). *The ITC international handbook of testing and assessment.* New York: Oxford University Press.

Levin, K., Willis, G. B., Forsyth, B. H., Norberg, A., Stapleton Kudela, M., Stark, D., y Thompson, F. E. (2009). Using cognitive interviews to evaluate the Spanish-language translation of a dietary questionnaire. *Survey Research Methods, 3* (1), 13-25.

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics: Development, analysis and application of psychological and educational tests*. The Hague, Netherlands: Eleven International.

Muñiz, J., Elosua, P., y Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema, 25,* 151-157.

Muñiz, J., Hernández, A., y Ponsoda, V. (2015). Nuevas directrices sobre el uso de los tests: investigación, control de calidad y seguridad. *Papeles del psicólogo, 36*, 161-173.

Padilla, J. L., y Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*, 136-144.

Pena, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development, 78,* 1255-1264.

Sireci, S. G., y Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2-3), 170-187.

Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W. J. G., Bouter, L. M., y de Vet, H. C. W. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. *Quality of Life Research, 21,* 651-657.

van de Vijver, F. J. R., y Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1* (2), 89-99.

van de Vijver, F. J. R., y Leung, K. (1997). Methods and data analysis of comparative research. En J. W. Berry, Y. H. Poortinga, y J. Pandey (Eds.), *Handbook of cross-cultural psychology (2nd ed.). Handbook of cross-cultural psychology, Vol 1: Theory and method* (pp. 257-300). Needham Heights, MA, US: Allyn y Bacon.

van de Vijver, F. J. R., y Leung, K. (2011). Equivalence and bias: A review of concepts, models and data analytic procedures. En D. Matsumoto y F. J. R. Van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 17-45). New York, NY: Cambridge University Press.

van de Vijver, F. J., y Poortinga, Y. H. (2004). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, y C. D. Spielberger (eds.). *Adapting educational and psychological tests for cross-cultural assessment* (pp. 51-76). Psychology Press.

van de Vijver, F. J. R., y Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée / European Review of Applied Psychology*, *54*(2), 119-135.

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., y Erikson, P. (2005). Principles of good practice for the translation and cultural adaptation process for Patient-Reported Outcomes (PRO) measures: Report of the ISPOR Task Force for translation and cultural adaptation. *Value in Health, 8,* 94-104.

Wolf, E. J., Harrington, K. M., Clark, S. L., y Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*, 913-934.

Resumen

Directrices ITC y criterios propuestos: Listado de verificación sobre el cumplimiento de los criterios

PRELIMINARES

DE DESARROLLO

Directrices	Criterios	No aplicable*	No aceptable	Aceptable	Excelente
DP1: Antes de comenzar con la adaptación hay que obtener los permisos pertinentes de quien ostente los derechos de propiedad intelectual del test.	DP1-1: Si la adaptación se considera la mejor opción, pedir los permisos correspondientes a los propietarios de los derechos de autor, incluso cuando el test solo se vaya a emplear con fines de investigación				
DP2: Evaluar si el grado de solapamiento en la definición y contenido del constructo medido mediante el test, así como en el contenido de los ítems, es suficiente para el uso (o usos) previsto(s) de las puntuaciones en las poblaciones de interés.	DP2-1: Proporcionar evidencia teórica y empírica de que el constructo de interés es relevante para la población diana a la que va dirigida la versión adaptada.				
	DP2-2: Considerar si el significado del constructo puede generalizarse a través de las culturas, y justificar que la traducción/adaptación del test es preferible a la creación de un test nuevo dirigido a la población diana.				
DP3. Minimizar la influencia de cualquier diferencia cultural o lingüística, que sea irrelevante para el uso previsto del test en las poblaciones de interés.	DP3-1: Si la adaptación se considera la mejor opción, evaluar las posibles diferencias culturales y lingüísticas antes de comenzar el proceso de adaptación. Estas diferencias deben ser consideradas en la versión adaptada, con el fin de prevenir sesgos y diseñar estudios que permitan controlar los potenciales sesgos.				
DD1: Asegurarse, mediante la selección de expertos cualificados, de que el proceso de traducción y adaptación tiene en cuenta las diferencias lingüísticas, psicológicas y culturales entre las poblaciones de interés.	D1-1: Formar un equipo multidisciplinar compuesto por: a) traductores profesionales para traducir el test de la lengua original a la lengua de adaptación (cuando sea necesaria la traducción) y que tengan cierto conocimiento de las culturas implicadas, b) expertos en el constructo a medir, c) expertos en las culturas implicadas, y d) expertos en la construcción de test. En algunos casos, un mismo miembro del equipo puede ser experto en más de uno de estos aspectos; por ejemplo, en las lenguas y culturas, en el constructo y las culturas, etc.				
DD2: Utilizar diseños y procedimientos racionales apropiados para asegurar la adecuación de la adaptación del test a la población a la que va dirigido.	DD2-1: Usar alguno de los diseños de traducción recomendados y justificar la elección. La traducción hacia adelante, hacia atrás (o retro-traducción) o la traducción simultánea son posibles alternativas, dependiendo del propósito de la traducción, del alcance del proyecto, del número de culturas implicadas y de si es o no necesario comparar las puntuaciones de personas que pertenecen a las distintas culturas implicadas. Para los diseños de traducción hacia adelante o hacia atrás, se requiere que se empleen al menos dos traductores (o equipos de traductores) independientes. Si un test se construye desde sus inicios con el fin de ser aplicado transculturalmente, es posible el desarrollo simultáneo/concurrente de las múltiples versiones del test desde el principio.				
	DD2-2: Contar con varios traductores que trabajan de forma independiente y constituir un comité de expertos que revisen y comparen las traducciones propuestas, con el fin de recoger sus opiniones sobre las versiones, resolver las discrepancias existentes, y proponer una versión consensuada.				
DD3: Ofrecer evidencias que garanticen que las instrucciones del test y el contenido de los ítems tienen un significado similar en todas las poblaciones a las que va dirigido el test.	DD3-1: Asegurar que las instrucciones del test son claras y comprensibles, empleando un lenguaje familiar para la población diana.				
	DD3-2: Asegurar que el contenido de los ítems resulta claro y se expresa con los mismos niveles de familiaridad y dificultad en la lengua original y la lengua diana. Los elementos lingüísticos que pudieran dificultar la comprensión de la versión traducida, como palabras con diferentes significados, dobles negaciones, etc. deben evitarse. Los elementos no verbales (imágenes o dibujos) deben estar contextualizados teniendo en cuenta la cultura de la población.				

Directrices	Criterios	No aplicable*	No aceptable	Aceptable	Excelente

DE DESARROLLO (Cont.)	DD4: Ofrecer evidencias que garanticen que el formato de los ítems, las escalas de respuesta, las reglas de corrección, las convenciones utilizadas, las formas de aplicación y demás aspectos son adecuados para todas las poblaciones de interés.	DD4-1: Intentar que el formato de los ítems, las opciones de respuesta, las rúbricas de puntuación si las hubiera, y el modo de aplicación del test sean similares en las distintas versiones.
		DD4-2: Asegurar que la población diana está suficientemente familiarizada con los procedimientos empleados (formato de los ítems, escalas de respuesta, rúbricas de puntuación (si las hubiera), y modo de aplicación) en el test adaptado.
	DD5: Recoger datos mediante estudios piloto sobre el test adaptado, y efectuar análisis de ítems y estudios de fiabilidad y validación que sirvan de base para llevar a cabo las revisiones necesarias y adoptar decisiones sobre la validez del test adaptado.	DD5-1: Comprobar la calidad psicométrica (análisis de ítems, fiabilidad y validez) de las puntuaciones del test adaptado en una muestra piloto de la población diana y, considerando los resultados, hacer las revisiones necesarias para mejorar la versión final del test. Las muestras del estudio piloto deben ser suficientemente grandes para realizar los análisis estadísticos correspondientes.
DE CONFIRMACIÓN	C1: Definir las características de la muestra que sean pertinentes para el uso del test, y seleccionar un tamaño de muestra suficiente que sea adecuado para las exigencias de los análisis empíricos” .	C1-1: Asegurar que la muestra diana sea lo suficientemente grande como para llevar a cabo los análisis estadísticos necesarios y para representar adecuadamente a la población.
		C1-2: Cuando el interés se centra en las comparaciones transculturales, asegurar que la muestra original y la muestra diana son comparables para todas las variables pertinentes, excepto el idioma y/o contexto cultural.
	C2: Ofrecer información empírica pertinente sobre la equivalencia del constructo, equivalencia del método y equivalencia entre los ítems en todas las poblaciones implicadas.	C2-1: Cuando haya interés en comparar la población original con la población diana, utilizar previamente procedimientos estadísticos para asegurar la equivalencia del constructo en las distintas poblaciones
		C2-2: Cuando haya interés en comparar la población original con la población diana, comprobar la equivalencia de método (características del instrumento, proceso de administración y características de la muestra).
		C2-3: Cuando haya interés en comparar las poblaciones objetivo y de referencia, evaluar el Funcionamiento Diferencial de los Ítems (DIF) entre los grupos culturales que se van a comparar utilizando procedimientos estadísticos adecuados al formato del ítem, el tamaño de la muestra y la dimensionalidad del test.
		C2-4: En caso de que el DIF se detecte en niveles significativos, llevar a cabo análisis para comprender las causas del DIF (por ejemplo, efectos lingüísticos o de método) entre las culturas.
	C3: Recoger información y evidencias que apoyen el uso de los baremos] la fiabilidad y la validez de la versión adaptada del test en las poblaciones implicadas.	C3-1: Asegurar que el tipo de indicadores de fiabilidad reportados son los adecuados para el tipo de test, utilizando análisis estadísticos y tamaños muestrales adecuados. Los valores obtenidos deben ser satisfactorios y acompañarse del error típico de medición.
		C3-2: Proporcionar evidencias de validez coherentes con el uso previsto de los resultados del test, utilizando análisis estadísticos y tamaños muestrales adecuados.
		C3-3: Asegurar y verificar que las normas son adecuadas para interpretar los resultados del test en la población diana.
	C4: Cuando se comparen puntaciones entre distintas versiones del test, emplear diseños de equiparación y análisis de datos adecuados que garanticen la comparabilidad.	C4-1: Cuando el objetivo es la evaluación intercultural o interlingüística y algunos ítems funcionan diferencialmente, se utilizan diseños de equiparación y procedimientos de análisis apropiados antes de la comparación.

	Directrices	Criterios	No aplicable*	No aceptable	Aceptable	Excelente
DE ADMINISTRACIÓN	A1: Preparar los materiales y las instrucciones para la aplicación de modo que minimicen cualquier diferencia cultural y lingüística que pueda ser debida a los procedimientos de aplicación y a los formatos de respuesta, y que puedan afectar a la validez de las inferencias derivadas de las puntuaciones.	A1-1: Para todos los materiales e instrucciones de aplicación del test, comprobar el cumplimiento de los requisitos establecidos en las directrices de desarrollo (TD3 a TD5). Para prevenir posibles problemas en la administración del test adaptado en la población diana, se ha de tener en cuenta la experiencia acumulada en la aplicación de la versión original del test en la población de origen.				
	A2: Especificar las condiciones de aplicación del test que deben seguirse en todas las poblaciones a las que va dirigido.	A2-1: Cuando se vayan a realizar comparaciones transculturales, asegurar que las condiciones de administración del test (modo de administración, limitación de tiempo, información sobre el propósito de la prueba, etc.) están estandarizadas a través de los grupos. En caso en que fuera necesario introducir cambios, se evaluar el posible impacto de las diferentes condiciones de administración del test.				
		A2-2: Asegurar que los entrevistadores o los administradores de los test tengan las credenciales necesarias según el tipo de test que se vaya a aplicar. Los administradores del test deben presentar una declaración firmada por la que se comprometen a llevar a cabo sus actividades de conformidad con el código de ética y los principios de la práctica profesional establecidos por las asociaciones y órganos profesionales nacionales competentes.				
PUNTIACIONES E INTERPRETACIÓN	SSI1: Interpretar las diferencias de las puntuaciones entre los grupos teniendo en cuenta la información demográfica pertinente.	SSI1-1: Cuando las comparaciones entre puntuaciones estén justificadas sobre la base del análisis de invarianza de la medida, interpretar las diferencias interculturales observadas (si las hubiera), según la información sistematizada y documentada en PC3-1 (G3, C4) en relación con la distancia cultural y lingüística. Para comprender las diferencias en las puntuaciones observadas, debe considerarse el papel de estas variables (por ejemplo, religiosidad, individualismo, diferentes tendencias de respuesta).				
	SSI2: Comparar las puntuaciones entre poblaciones únicamente en el nivel de invarianza establecida para la escala de puntuación utilizada en las comparaciones.	SSI2-1: Antes de comparar las puntuaciones individuales de personas pertenecientes a diferentes culturas, y/o las puntuaciones promedio de grupos de diferentes culturas, se debe asegurar que la equivalencia métrica (i.e. la ausencia de DIF) ha sido evaluada y se cumple, al menos para un número relevante de ítems.				
DOCUMENTACIÓN	Doc1: Proporcionar documentación técnica que recoja cualquier cambio en el test adaptado, incluyendo la información y las evidencias sobre la equivalencia entre las versiones adaptadas.	Doc1-1: Crear una serie de documentos y ponerlos a disposición de las personas y sectores interesados, proporcionando información de las 8 cuestiones listadas en el apartado correspondiente a este criterio.				
	Doc2: Proporcionar documentación a los usuarios con el fin de garantizar un uso correcto del test adaptado en la población a la que va dirigido.	Doc2-1: Asegurar que los materiales y la documentación que acompañan al test (p.e., el manual del test) sean claros (instrucciones, descripción del ámbito de aplicación, ejemplos prácticos de su uso, etc.) para garantizar que el test sea adecuado para la población a la que se dirige, que la administración del test esté estandarizada y que las puntuaciones se puedan interpretar adecuadamente (véanse las secciones de aplicación y puntuación).				

* Si uno o más de los criterios no son aplicables (por ejemplo, SSI2-1, cuando el propósito de la adaptación no es comparar puntuaciones de personas pertenecientes a distintas culturas, o C2-4, cuando el análisis del DIF indica que no hay ítems con funcionamiento diferencial) esto debe hacerse explicito. Si no se dispone de información suficiente para juzgar si un criterio es aceptable o no (a pesar de que esta información es relevante considerando el propósito de la adaptación), la valoración será “No aceptable”.