LA ADAPTACIÓN DE LOS TESTS EN ESTUDIOS COMPARATIVOS INTERCULTURALES

TESTS ADAPTATION IN CROSS-CULTURAL COMPARATIVE STUDIES

BARBERO GARCÍA, Mª. I.; VILA ABAD, E. Y HOLGADO TELLO, F.P.[1]

[1]Dpto. Metodología de las Ciencias del Comportamiento
Faculta de Psicología
UNED

Resumen

La creciente globalización que se está dando a nivel mundial, el problema de la emigración y sus necesidades de adaptación, el intercambio de estudiantes entre distintos países fortalecido, además, por la creación de un Espacio Europeo de Educación Superior (EEES), y el creciente interés que actualmente tienen muchos países por establecer unos estándares educativos internacionales que permitan comparar los progresos que, cada uno de ellos, han hecho en relación con los demás, son algunos de los factores que han potenciado la preocupación porque los instrumentos de medición utilizados sean los adecuados y cumplan los requisitos psicométricos necesarios para su utilización en estudios comparativos interculturales. Las diferencias encontradas en el rendimiento obtenido por los estudiantes de los distintos países, han de deberse a diferencias reales y no a otra serie de factores (diferencias culturales, idiomáticas, etc.) que invaliden la interpretación de los resultados.

El objetivo que nos proponemos con este trabajo es analizar brevemente las fases principales que se deben cubrir a la hora de llevar a cabo la adaptación de un test a distintas culturas o idiomas, siguiendo las pautas marcadas por la Comisión Internacional de Tests, con el fin de conseguir esos requisitos psicométricos que antes se comentaban, y hacerlo de una manera sencilla, que pueda resultar útil a todas aquellas personas que, en alguna medida, estén

Abstract

The globalization phenomenon, the emigration and its needs of adaptation, the students mobility strengthened by the European Higher Education convergence, and the interest of the countries in stabilising international psychological and educative standards that make possible comparative studies, are some of the factors that has been promoted the interest about the instruments used in cross-cultural studies. The differences found should be based on real differences and not on other factors (culture, language, etc.) that could contaminate the interpretation of the results. Considering this problem the main objective of this paper is to analyse the phases established by the International Test Commission in test adaptation and translation. We try to show these phases in a simple way in order to provide a practical guide to the people related to psychological and educational evaluation.

relacionadas con el problema de la adaptación de instrumentos para la evaluación psicológica y/o educativa.

Palabras Clave

Adaptación de tests, estudios transculturales.

Key Words

Tests adaptations, cross-cultural studies.

Introducción

Teniendo en cuenta que los tests son los instrumentos más utilizados para llevar a cabo los estudios comparativos interculturales, la calidad científica de los mismos, así como su correcta o incorrecta utilización, es uno de los problemas que se deben abordar. Lo primero que hemos de tener en cuenta es que los tests que han sido validados en un país no pueden ser directamente utilizados en cualquier otro; es necesario llevar a cabo un nuevo proceso de validación para adaptarlo a la nueva situación teniendo en cuenta las características propias de cada uno de los países a los que se va a adaptar y, finalmente, elaborar baremos nuevos.

La importancia del problema se puso ya de manifiesto cuando en 1985 la American Educational Research Association (AERA) junto con la American Psychological Association (APA) y el National Council on Measurement in Education (NCME) publicaron los *Standards for Educational and Psychological Testing* que proporcionaron un marco teórico para ayudar en el proceso de adaptación de los tests y, aunque hoy en día han sido modificados y completados con las nuevas directrices elaboradas por la Comisión Internacional de Tests (ITC), sirvieron para alertar acerca de las posibles fuentes de error que pueden surgir durante el proceso de adaptación y que hay que tener muy en cuenta.

El Colegio Oficial de Psicólogos (COP) participó en la elaboración de las directrices cuya finalidad es orientar a los profesionales en la difícil tarea de adaptar los tests de unos países a otros (Hambleton y Bollwark, 1991; Hambleton, 1993, 1994,1996; Hambleton, Merenda y Spielberger, 2005; Muñiz,1996; Van der Vijver y Hambleton, 1996; Van der Vijver y Poortinga, 1991). En la página Web del COP (http://www.cop.es/tests/) se puede encontrar el documento completo que recoge las 22 directrices junto con otra serie de documentos importantes.

Fases del proceso de adaptación

Estas directrices han sido divididas en cuatro grandes áreas que hacen alusión a las distintas fases del proceso de adaptación en las que se pueden producir errores y que, tal y como hemos apuntado, hay que tratar de minimizar: 1) *contexto;* 2) *construcción y adaptación,* 3) *aplicación* y 4) *interpretación de las puntuaciones*.

Contexto

Teniendo en cuenta que la mayoría de los constructos utilizados en Psicología tienen una fuerte dependencia de aspectos culturales, lo primero que hay que analizar es la equivalencia del constructo en las distintas culturas en las que se va a aplicar el test para poder comparar los resultados obtenidos. Hay veces que un mismo constructo tiene un significado distinto, o puede tener distinta interpretación, en las diferentes culturas. Por ejemplo, no todos los países tienen el mismo concepto de lo que se entiende por inteligencia, calidad de vida, etc. Mientras que en Occidente, el constructo inteligencia está asociado a conductas eficientes y a la rapidez de actuación, en algunas culturas del Este se asocia a conductas reflexivas y reposadas (Lonner, 1990). En este caso, al no ser equivalentes los constructos a medir carecería de sentido todo el proceso de adaptación puesto que se estarán midiendo constructos distintos.

Cuando se van a llevar a cabo estudios interculturales los investigadores pueden evaluar la equivalencia del constructo bien antes de llevar a cabo el estudio empírico propiamente dicho y, por lo tanto, sin disponer de los datos, o bien después del estudio de campo. En cualquier caso ambas formas de evaluación no sólo no son incompatibles sino que se complementan ya que la información que se obtiene en ambas es distinta.

Cuando no se dispone de los datos y se quiere examinar la equivalencia del constructo en distintas culturas o idiomas para llevar a cabo la adaptación de distintas versiones del test, se puede utilizar el juicio de expertos que representen las diferentes culturas e idiomas. Si hay consistencia en los juicios de los expertos se contará con una evidencia preliminar acerca de la equivalencia del constructo.

Una vez que se ha conseguido esta evidencia preliminar, se puede pasar a la fase siguiente del proceso, la construcción de las distintas versiones del test, con las que, una vez aplicadas, se intentaran medir las conductas concretas que han sido seleccionadas como manifestaciones del constructo.

Si ya se ha llevado a cabo la aplicación del test en sus distintas versiones y se cuenta con las respuestas de los sujetos, se pueden utilizar distintas técnicas estadísticas para abordar el problema de la equivalencia del constructo en las distintas versiones; entre las más utilizadas están el análisis factorial, tanto exploratorio (AFE) como confirmatorio (AFC), y el escalamiento multidimensional (EM).

a) *Respecto al análisis factorial exploratorio*, van der Vijver y Poortinga (1991) y Poortinga (1991) consideran que, a pesar de algunas críticas, es la técnica estadística más utilizada para evaluar cuándo un constructo que tiene una determinada estructura en una cultura manifiesta las mismas características en otra. La utilización de esta técnica implica llevar a cabo el análisis por separado en cada grupo cultural (o lingüístico) y, a continuación, observar las matrices de pesos factoriales resultantes para evaluar su consistencia en los distintos grupos. Aunque no hay un acuerdo generalizado para decidir cuando dos estructuras factoriales se pueden considerar equivalentes, dado que se trata de una tarea subjetiva, se han encontrado algunos índices como el de Burt (1948) y Tucker (1951) que pueden facilitar la tarea.

b) *La utilización del análisis factorial confirmatorio* implica hipotetizar «a priori»cuál va a ser la estructura factorial del test y a partir de las puntuaciones de los sujetos se verificará la viabilidad de dicha estructura (Byrne, 1998, 2001, 2003). Esta estructura hipotetizada se incorpora en un modelo de ecuaciones estructurales y se contrasta a través de los grupos bajo la hipótesis de que es la misma en todos ellos. Si la matriz de pesos factoriales es equivalente en todos los grupos se tendrá evidencia de la equivalencia del constructo. En estudios interculturales se utiliza el AFC para evaluar cuándo la estructura factorial de la versión original de un test es consistente a través de las siguientes versiones que han sido traducidas a otros idiomas (Brown y Marcoulides, 1996; Reise, Widaman y Pugh, 1993; Robie y Ryan, 1996; Sireci, Bastari y Allalouf, 1998). A diferencia de lo que ocurría con el AFE, cuando se utiliza el AFC se pueden analizar simultáneamente los datos provenientes de todos los grupos.

Esta técnica puede ser problemática si los ítems se han puntuado de forma dicotómica puesto que los modelos subyacentes son modelos lineales y las relaciones entre los ítems dicotómicos no lo son; sin embargo, hay algunas soluciones, por ejemplo en lugar de utilizar las correlaciones de Pearson utilizar correlaciones policóricas (en España, Barbero, Holgado, Vila y Chacón, 2007 entre otros) o agrupar los ítems en distintos bloques antes del análisis.

c) *El escalamiento multidimensional* no requiere que se especifique a priori la estructura del test como ocurría en el AFC, pero al igual que en esta técnica se pueden analizar simultáneamente los datos provenientes de todos los grupos, lo cual constituye una ventaja. El escalamiento multidimensional proporciona información acerca de la estructura subyacente a los datos y permite evaluar cuándo esa estructura es consistente a través de los grupos de interés. Por otra parte no requiere un modelo lineal para derivar la estructura subyacente a los datos.

Construcción y adaptación del test

Aunque, tal y como se ha comentado anteriormente, a la hora de construir una segunda

forma de un test para su utilización en otra cultura o en otro idioma es necesario conseguir que mida el mismo constructo, hay que tener en cuenta las peculiaridades de la población a la que se va a aplicar; por lo tanto, es fundamental que aquellas personas que van a llevar a cabo la adaptación, además de tener conocimientos acerca del proceso de construcción de un test, conozcan los idiomas correspondientes a las distintas poblaciones y las peculiaridades culturales de cada una de ellas. Es fácil comprender la dificultad que entraña todo este proceso si se quiere hacer bien.

Una técnica habitual, cuando se trata de tener que adaptar un test de un idioma a otro, es que un equipo haga una traducción del test del idioma original al de la nueva población y, a continuación, otro equipo diferente vuelva a traducir el test al idioma original (*traducción inversa*). A medida que la coincidencia entre las traducciones sea mayor tendremos una mayor seguridad de que el proceso ha sido llevado a cabo de forma correcta. A pesar de que es una técnica muy utilizada no está libre de inconvenientes que han tratado de paliarse con otros procedimientos que permitirán, además de establecer la equivalencia de los ítems en cada una de las versiones del test, la posterior equiparación de las puntuaciones obtenidas en cada uno de los grupos para su comparación.

Los tres procedimientos más utilizados son: a) Diseño de grupos separados monolingües, en los que se administra la versión original del test y la versión adaptada, de forma separada, a los respectivos grupos lingüísticos; b) Diseño de grupos bilingües en los que a un grupo de sujetos bilingües se les aplican las dos versiones del test, la original y la adaptada y c) Diseño de un solo grupo de monolingües en los que a un grupo de monolingües en el idioma original se les aplica la versión original del test y la traducción inversa del mismo.

Una mayor información de las ventajas e inconvenientes de cada uno de estos diseños se puede encontrar en Hambleton (1996).

Estos mismos diseños se pueden utilizar, con las consiguientes variaciones, cuando en lugar de tratarse de adaptaciones de un test a distintos idiomas, se hacen a distintas culturas. Aún dentro de un mismo idioma hay términos que tienen connotaciones diferentes en distintas poblaciones y esto hay que tenerlo en cuenta a la hora de la adaptación del test, esta tarea se puede llevar a cabo por expertos.

Una vez hecha la traducción es necesario comprobar las propiedades psicométricas de la nueva forma del test, por lo tanto hay que comprobar su fiabilidad, validez, y llevar a cabo el proceso de estandarización. A pesar de la dificultad y trabajo que conlleva la elaboración de nuevos baremos, tal y como señala Muñiz (1996), bajo ningún concepto se deben utilizar los baremos elaborados en otra población ya que uno de los aspectos clave que hay que comprobar es que los ítems no estén sesgados contra un determinado grupo y, por lo tanto, no muestren un funcionamiento diferencial.

Impacto, sesgo y funcionamiento diferencial de los ítems

Hoy en día existen numerosas técnicas estadísticas que permiten el estudio del funcionamiento diferencial de los ítems. Pero, antes de pasar a su enumeración hemos de diferenciar entre tres importantes términos: *impacto, funcionamiento diferencial y sesgo del ítem.* Tal y como señala Fidalgo (1996), cuando se dice que un ítem presenta funcionamiento diferencial, lo único que se indica es que muestra diferentes propiedades estadísticas en diferentes grupos; ahora bien, aunque esto ocurre siempre que un ítem presenta funcionamiento diferencial, no siempre podemos hablar de que el ítem muestra funcionamiento diferencial, también puede deberse a que hay impacto. El término *impacto* hace referencia a una diferencia significativa de un grupo en un ítem, por ejemplo, un grupo responde correctamente en mayor proporción que otro grupo a un ítem. Ahora bien esa diferencia entre los grupos es debida a una diferencia real en la variable medida. Por el contrario, si los sujetos que tienen el mismo nivel en la variable medida difieren en la probabilidad de responder correctamente a un ítem, entonces el ítem muestra *funcionamiento diferencial (FDI)* y por lo tanto está *sesgado* frente a uno de los grupos de interés. El análisis del funcionamiento diferencial del ítem pondrá de manifiesto a qué es debido el sesgo. Para ello, los sujetos de los dos grupos de interés se emparejan en rela-

ción con la habilidad que está siendo medida. Una vez emparejados, los sujetos de distintos grupos que tienen el mismo nivel de habilidad deberán responder de una forma similar al ítem. Si esto no ocurre se dice que el ítem muestra un funcionamiento diferencial a través de los grupos. El análisis del impacto del ítem o del funcionamiento diferencial es de naturaleza estadística, el análisis del sesgo del ítem es esencialmente cualitativo. Se dice que un ítem está *sesgado* frente a un grupo cuando los sujetos que pertenecen a ese grupo tienen un menor rendimiento que los sujetos con igual nivel de habilidad pero que pertenecen a otro grupo, tomado como grupo de referencia, *y la razón de este menor rendimiento es irrelevante (variables ajenas) para el constructo que mide el test*. Ahora bien, para que se pueda hablar de que existe sesgo del ítem es necesario identificar alguna característica del ítem que esté perjudicando a uno o más grupos (por ejemplo cuando el ítem aborda un concepto que es más familiar a un grupo que a otro y sin embargo ese concepto no es central en la habilidad que se está midiendo).

Una vez que mediante las técnicas estadísticas adecuadas se detectan aquellos ítems que muestran funcionamiento diferencial, mediante un análisis cualitativo se intentarán explicar a qué son debidas las diferencias observadas y, una vez que se ha encontrado la explicación para estas diferencias y se observa que son diferencias no relacionadas con el objetivo del test, al ítem en cuestión se le añade la etiqueta de «*ítem sesgado*».

La problemática de los ítems señalados con FDI puede ser debida a una deficiente traducción, o al uso de un término, situación o expresión que son desconocidos o poco familiares para una de las poblaciones, aunque existan otras posibles causas. Quizás la habilidad que mide el ítem no forma parte del repertorio del idioma de la población objetivo, o tal vez el formato del ítem no les resulta familiar. Es importante determinar la razón de la diferencia porque influye en la decisión final acerca de qué hacer con el ítem.

Evaluación del funcionamiento diferencial de los ítems

Una vez evaluados los ítems del test, y constatado que «a priori» son adecuados para las poblaciones a las que va dirigido, es necesario aportar datos estadísticos de su equivalencia. Este es uno de los aspectos fundamentales en el proceso de validación de un test para su utilización en dos o más poblaciones cultural o idiomáticamente distintas. Básicamente, para que se pueda mantener que existe equivalencia de dos poblaciones en un test se requiere que haya datos que avalen que cuando los miembros de ambas poblaciones tienen el mismo nivel en la variable medida, ambos puntúan de forma equivalente en cada ítem.

Una vez expuestas las diferencias entre impacto, sesgo y funcionamiento diferencial de los ítems, pasamos a enumerar algunas de los procedimientos más utilizados hoy en día para llevar a cabo los análisis requeridos del FDI.

Básicamente se pueden diferenciar tres tipos de procedimientos:1) procedimientos basados en la Teoría de Respuesta a los Ítems, 2) procedimientos de Mantel-Haenszel (MH) y extensiones (Hambleton et al, 1993; Holland y Thayer, 1988; Holland y Wainer, 1993 y 3) procedimientos de regresión logística (Swaminathan y Rogers, 1990). En España han sido muchos los trabajos realizados para evaluar el FDI utilizando distintos procedimientos: (Barbero y Prieto, 1997; Prieto, Barbero y San Luis, 1999; Barbero, Prieto y San Luis, 2000; Barbero, Suárez, Prieto y San Luis, 2002, Elosúa, y López, 1999; Elosúa, López y Torres, 1999; Elosúa, López y Egaña, 2000; Gómez e Hidalgo, 1997; Hidalgo y Gómez, 1999, 2003, 2006a y b, 2007a b y c; Prieto, Barbero y San Luis, 1997, entre otros). Todas estas metodologías son «condicionales» en el sentido de que las comparaciones se hacen entre grupos de personas que se asumen igualados en la aptitud medida por el test. En los procedimientos basados en la Teoría de Respuesta al ítem los sujetos se igualan utilizando puntuaciones estimadas de aptitud (estimadas utilizando los patrones de las puntuaciones de los ítems). En los otros dos procedimientos se utiliza la puntuación total en el test para igualar a los sujetos (o una puntuación ajustada tras eliminar los ítems dudosos). Los tres tipos de procedimientos pueden producir resultados fiables y válidos, siempre que el tamaño de las muestras sea el apropiado, se apliquen correctamente, y los resultados se interpreten bien.

Para el caso del procedimiento de Mantel-Haenszel y la regresión logística se necesitan muestras de alrededor de 200 sujetos para cada población. En general, si se utilizan los procedimientos basados en la Teoría de Respuesta a los Ítems se necesitan muestras considerablemente más numerosas, si bien el modelo de Rasch requiere tamaños muestrales equivalentes a los otros dos procedimientos.

Cuando se encuentran ítems que no funcionan de forma equivalente en los distintos tests adaptados puede deberse o bien a que están mal adaptados, o a que estando bien adaptados son culturalmente inapropiadas (Hulin,1987), en ambos casos no pueden utilizarse en estudios comparativos puesto que proporcionan información diferente en las distintas poblaciones que se están comparando Las preguntas deficientemente adaptadas pueden revisarse (si se pretenden utilizar de nuevo) o eliminarlas. Sin embargo, preguntas bien adaptadas, aunque consideradas no equivalentes (o culturalmente inapropiadas) pueden proporcionar a menudo información adicional útil acerca de la población específica que se está comparando. La identificación de la fuente de no equivalencia de estas preguntas puede arrojar luz sobre las poblaciones culturales/idiomáticas respectivas y mejorar la comprensión de esa población (Ellis, 1991).

Aplicación

Se deben cuidar al máximo aspectos tales como las instrucciones que se den antes de la aplicación y las interacciones entre el aplicador y los examinados ya que pueden influir en la fiabilidad y validez del test. Los aplicadores deberán seleccionarse entre las personas pertenecientes a la población a la que se aplica el test; de esta manera estarán familiarizados con los distintos matices propios de dicha cultura. Por otra parte es conveniente que reciban un cierto entrenamiento sobre la forma de aplicación, haciéndoles ver la importancia que tiene el seguir al pié de la letra las instrucciones de aplicación.

En general, en la medida que las diferencias culturales o idiomáticas entre las poblaciones a las que se va a aplicar el test sean mayores, los problemas derivados de la aplicación del test aumentarán, de ahí que se necesite conocer bien la cultura e idioma de los grupos objetivo para tratar de paliar estos problemas. Una buena práctica consiste en proporcionar una lista de los problemas que ocurren con mayor frecuencia y los factores que amenazan la validez, proponiendo una serie de medidas a tomar.

En algunos casos los constructores pueden disponer de datos obtenidos en minorías culturales o aplicaciones interculturales ya realizadas. Toda esta información relevante debe ser recogida en el manual del test.

Las características del administrador, como el género, edad e incluso la forma de vestir pueden influir en el resultado de la medición, especialmente en los tests de aplicación individual. Cuando un test nuevo o adaptado se aplica a un determinado grupo es fácil, posiblemente con la ayuda de informadores locales, poner de manifiesto qué características de los aplicadores pueden poner en peligro la validez de los resultados del test, y así se pueden tomar las medidas oportunas (como por ejemplo realizar un estudio piloto). En el caso de que el aplicador y los examinados provengan de distintas tradiciones culturales, hay que analizar con cuidado el impacto negativo de los aplicadores y tomar las medidas necesarias para minimizar cualquier problema que se identifique.

Interpretación de las puntuaciones

Para poder interpretar adecuadamente los resultados obtenidos al aplicar un test, tanto si es un test original como cualquier adaptación, es necesario tener una cierta formación psicométrica; pero, en este último caso, existe además el peligro de que al comparar los resultados obtenidos en distintos grupos o países se intente hacer algún tipo de clasificación de los sujetos. Una cosa es analizar las semejanzas y diferencias entre los grupos y otra establecer cualquier tipo de clasificación, puesto que es prácticamente imposible encontrar dos comunidades que sean equiparables completamente en aspectos como, por ejemplo, la motivación a la hora de hacer las pruebas, valores culturales, nivel de vida, etc. (Muñiz, 1996). Para una mejor interpretación de los resultados, el manual del test deberá incluir una información exhaustiva de todo el proceso de adaptación.

Para muchos especialistas en medición, así como para los usuarios de los tests adaptados, la información acerca del proceso de adaptación del test puede proporcionar una gran ayuda para saber si es adecuado utilizar el test dentro de un contexto determinado. La documentación debería incluir una exposición detallada, paso a paso, del procedimiento seguido, incluyendo el diseño utilizado, los métodos usados para evaluar la equivalencia entre las versiones adaptadas, identificación, selección y labor de los traductores, las razones y justificaciones del uso y la inclusión de los ítems, así como información acerca de aquellos ítems que fueron modificados o excluidos, algunos de los problemas más importantes encontrados y cómo se resolvieron, todos los aspectos relativos a la aplicación de los tests, incluyendo la selección y entrenamiento de los aplicadores, y la interpretación de los resultados.

En la práctica, uno de los errores habituales es llevar a cabo un proceso de adaptación no del todo riguroso y después interpretar las diferencias entre las puntuaciones de las muestras como si fuesen auténticas. Esta negligencia a la hora de tratar los problemas de adaptación de los tests y validarlos en las culturas en las que se utilizan, ha debilitado seriamente los resultados de muchos estudios interculturales. Hay que ser muy cautelosos a la hora de interpretar los resultados obtenidos en diferentes poblaciones.

Hay veces que es posible poner en una escala común las puntuaciones obtenidas a partir de diferentes versiones de un test. Mediante la utilización de grandes muestras y potentes modelos estadísticos, como los de la Teoría de Respuesta al Ítem (Hambleton y col. 1991) se pueden llevar a cabo complejos sistemas de equiparación entre las puntuaciones de distintas versiones de un test, siempre, claro está, que el constructo sea razonablemente equivalente y se disponga de los datos correspondientes. Ahora bien, las puntuaciones de diferentes versiones idiomáticas de un test no siempre están equiparadas, en cuyo caso no pueden compararse directamente. Aún así, se pueden hacer cierto tipo de comparaciones, por ejemplo comparar a validez predictiva de distintas versiones de un test, pero los investigadores deberán limitar sus interpretaciones a aquellas situaciones para las que se dispone de datos fehacientes sobre la validez.

Tal y como apunta Muñiz (1998), el problema de la equivalencia de las puntuaciones, obtenidas en distintas versiones de un mismo test para su comparación no ha sido un problema que interesara demasiado a los psicómetras hasta la década de los ochenta en la que Lord (1980) dedica un capítulo y los *Standards for Educational and Psychological Testing* (1985*)* incluyen varios estándares relacionados con el tema.

Las causas del interés se deben, fundamentalmente, a la utilización de los tests para llevar a cabo estudios a gran escala, lo que obliga a tener diversas formas de un mismo test con el consiguiente problema de comparar y equiparar las puntuaciones que obtienen en ellos los sujetos, y la exigencia de la sociedad a los constructores de justificar y explicar públicamente los sistemas utilizados para la equiparación de las puntuaciones.

Angoff (1984) define la equiparación como un proceso que consiste en desarrollar un sistema que permita convertir las unidades de un test a las de otro. Se trata de establecer una correspondencia entre las puntuaciones de varios tests de forma que las puntuaciones, una vez realizada la equiparación, fueran totalmente equivalentes e intercambiables. Por lo tanto, al final del proceso de equiparación se obtendría una ecuación de equivalencia o una tabla de conversión. Para poder derivar la ecuación de equiparación y conseguir puntuaciones realmente equivalentes es necesario satisfacer una serie de requisitos tal y como plantea Angoff (1984) y recoge Navas (1996, pág. 300).

- *Todos los tests deben medir el mismo constructo* (esto ya se ha evaluado con anterioridad)
- *La ecuación de equivalencia debe de ser independiente del grupo de sujetos utilizado en su construcción.* A este requisito se le conoce como «invarianza de la población».
- *La equiparación debe de ser simétrica, es decir, de modo que sea lo mismo tomar como base las puntuaciones obtenidas en un test o las obtenidas en el otro a la hora de establecer la equivalencia.*
- *Una vez realizada la equiparación las puntuaciones de los tests deben de ser total-*

mente intercambiables, es decir, después de la equiparación deben de ser idénticas las distribuciones condicionales de las puntuaciones de cada test, dado un determinado nivel en el rasgo o característica medida.

Finalmente, será necesario evaluar el proceso de equiparación para obtener información acerca de hasta qué punto se ha hecho de forma correcta. Se trata de una fase crucial a pesar de que es obviada en muchos trabajos, pues permite comprobar hasta qué punto se ha conseguido que las puntuaciones estén en la misma escala.

Discusión

En este trabajo nos hemos planteado un doble objetivo, por una parte, ofrecer una guía práctica que facilitara la tarea a todos aquellos profesionales que en algún momento se encuentren inmersos en un proceso de adaptación de tests a distintas culturas o idiomas y, por otra, alertar acerca de la dificultad del proceso y de la necesidad de tomar todas las medidas necesarias para que, realmente, se lleve a cabo con las garantías que debe tener todo trabajo científico. De esta manera las inferencias que se hagan a partir de los datos obtenidos serán correctas.

Tomando como guía las cuatro grandes áreas en que fueron divididas las directrices elaboradas por la Comisión Internacional de Tests (ITC), se han ido comentando las principales fuentes de error que se pueden originar en cada una de ellas y las posibles acciones a realizar para minimizarlos.

Referencias bibliográficas

American Educational Research Association (AERA), American Psychological Association (APA) y National Council on Measurement in Education (NCME) (1985). *Standards for Educational and Psychological Testing*. Washington DC: American Psychological Association.

American Educational Research Association (AERA), American Psychological Association (APA) y National Council on Measurement in Education (NCME) (1999). *Standards for Educational and Psychological Testing*. Washington DC: American Psychological Association.

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

Barbero, M.I. y Prieto, P. (1997). Evaluación del rendimiento en Ciencias de los niños y niñas de 13 años en las distintas comunidades Autónomas: Impacto o sesgo. *Psicothema*, 9(2), pp. 433- 440.

Barbero, M.I., Prieto, P. y San Luis, C. (2000). Procedimientos para la detección de FDI tanto en ítems politómicos como dicotómicos. *Psicothema*, vol 12 (2), 69-73.

Barbero, M.I., Suárez, J.C., Prieto, P. y San Luis, C. (2002). Invarianza de los parámetros en el modelo de Rasch. Un estudio de simulación. *Revista de Metodología de las Ciencias del Comportamiento*. Suplemento (1) 73-78.

Barbero, M.I., Holgado, F.P., Vila, E. y Chacón, S. (2007). Actitudes, hábitos de estudio y rendimiento en matemáticas: Diferencias por género. *Psicothema*, 10(3), pp. 413- 421.

Brown, R. y Marcoulides, G.A. (1996). Across- cultural comparison to the Brown locus of control scale. *Educational and Psychological Measurement,* 56, 858-863.

Burt, C. (1948). The factorial study of temperamental traits.*British Journal of Psychology, statistical section,* 1, 178-203.

Byrne, B. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS*. Mahwah, NJ: Lawrence Earlbaum Associates.

Byrne, B. (2001). Structural equation modeling with AMOS, EQS, and LISREL. Comparative approaches to testing for the factorial validity of a measuring instrument. *International Journal of Testing*. 1, 55-86.

Byrne, B. (2003). Confirmatory factor analysis. En R. Fernández- Ballesteros (Ed.), *Enciclopedia of Psychological Assessment* (vol 1). Thousand Oaks, CA: Sage.

Ellis, B.B. (1991). Item response theory: A tool for assessing the equivalence of translated tests. *Bul-*

letin of the International Test Commission, 18, 33-51.

Elosúa, P y López, A. (1999). Funcionamiento diferencial de los ítems y sesgo en la adaptación de dos pruebas verbales. *Psicológica,* 20, 23-40.

Elosúa, P; López, A. y Torres, E. (1999). Adaptación al euskera de una prueba de inteligencia verbal. *Psicothema,* 11 (1)151-161.

Elosúa, P.; López, A. y Egaña, J. (2000). Fuentes potenciales de sesgo en una prueba de aptitud numérica. *Psicothema,* 12, (3), 376- 382.

Fidalgo, A.M. (1996). Funcionamiento diferencial de los ítems. En J. Muñiz (Coord.), *Psicometría*,(pp. 371-455), Madrid: Universitas

Gómez, J. e Hidalgo, M.D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: Una revisión metodológica. *Anuario de Psicología,* 74, 3-32.

Hambleton, R.K. y Bollwark, J. (1991). Adapting tests for use in different cultures: Technical issues and methods. *Bulletin of the International Test Commission,* 18, 3-32.

Hambleton, R.K. (1993). Translating achievement tests for use in cross- national studies. *European Journal os Psychological Assessment,* 9, 57-68.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, pp.229-240.

Hambleton, R.K. (1996). Adaptación de tests para su uso en diferentes idiomas y culturas: fuentes de error, posibles soluciones y directrices prácticas. En J. Muñiz (Coord.), *Psicometría* (pp. 203- 238). Madrid: Universitas.

Hambleton, R.K. Merenda, P.F. y Spielberger, Ch.D. (2005). *Adapting Educational and Ppsychological Tests for Cross-Cultural Assessment*. Mahwah, NJ: Lawrence Earlbaum Associates.

Hidalgo, M.D. y Gómez, J. (1999). Técnicas de detección de funcionamiento diferencial en ítems politómicos. *Metodología de las Ciencias del Comportamiento,* 1, 39-60.

Hidalgo, M.D. y Gómez, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment,* 19, 1-11.

Hidalgo, M.D. y Gómez, J. (2006 a) . Nonuniform DIF detection using discriminant logistic analysis and multinomial logistic regression: a comparison for polytomous items. *Quality & Quantity,* 40, 805- 823.

Hidalgo, M.D. y Gómez, J. (2006 b). DIF detection in small samples: logistic regression analysis and effect size measure. *Paper presented at the 5th Conference of the International Test Commission*, Bruselas, Bélgica, Julio 8-9.

Hidalgo, M.D. y Gómez, J. (2007 a). Almost forty years of Differential Item Functioning: efficacy and utility. Comunicación presentada en el Symposium « Methodological issues in cross- cultural Psychology» en el Xth European Congress of Psychology, Praga, 3-6 de Julio.

Hidalgo, M.D. y Gómez, J. (2007 b). Education Measurement: Differential Item Functioning. En *International Encyclopedia of Education*. 3rd. Ed.

Hidalgo, M.D. y Gómez, J (2007 c). DIF detection in small simples using logistic regression analysis with effect size measure: Manuscript submitted for publication.

Holland, P.W. y Wainer, H. (1993) *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P.W.y Thayer , D.T. (1988). Differential item functioning and the Mantel-Haenszel procedure. En H. Wainer y H.I. Braun (Eds.), *Test validity* (pp. 129-145). Hilldale, NJ: Lawrence Erlbaum Associates.

Hulin, C.L. (1987). A psychometric theory of evaluations of item and scale translation: Fidelity across languages. *Journal of Cross-Cultural Psychology,* 18, 115-142.

Lonner, W.J. (1990) An overwiew of cross-cultural testing and assessment. En R.W. Brislin (Ed.), *Applied cross-cultural psychology (* vol. 14, pp. 56-76). Newbury Park, CA: Sage

Lord, F.M. (1980) *Applications of item response theory to practical testing problems* Hillsdale NJ: Lawrence Erlbaum Associates.

Mantel, N. y Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 19-48.

Muñiz, J. (1996). *Psicometría.* Madrid: Universitas.

Muñiz, J. (1998). *Teoría clásica de los tests*.. Madrid: Pirámide.

Navas, M.J. (1996). Equiparación de puntuaciones. En J. Muñiz (Coord.), *Psicometría*, (pp. 293- 369), Madrid: Universitas.

Poortinga, Y.H. (1991). Conceptual implications of item bias. En P.L. Dann, S.H. Irvine y J.M. Collins (Eds.) *Advances in computer-based human assessment (*pp. 279-290). Dordrecht, Netherlands: Kluwer Academic.

Prieto, P., Barbero, M.I. y San Luis, C. (1997) Identification of nonuniform dif. A comparison of Mantel – Haenszel and IRT analysis procedures. *Educational and Psychological Measurement*. Vol 57 (4), 559 – 568.

Prieto, P. Barbero, M.I. y San Luís, C.(1999). Detección del funcionamiento diferencial de los ítems en una prueba de Ciencias. *Psicothema,* 11(3), 691-697.

Reise, S.P., Widaman, K.F. y Pugh, R.H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin,* 114, U552-U566.

Robie, C. y Ryan, A.M. (1996). Structural equivalente of a measure of cross-cultural adjustment. *Educational and Psychological Measurement*, 56, 514-521.

Sireci, S.G., Bastari, B. y Allalouf, A. (August, 1998). *Evaluating construct equivalence across adapted tests*. Invited paper presented at the meeting of the American Psychological Association, San Francisco.

Swaminathan, H. y Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement,* 27, 361- 370.

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

van der Vijver, F.J. y Poortinga,Y.H. (1991). Testing across cultures. En R.K. Hambleton y J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277- 308). Dordrecht, Netherlands: Kluwer Academic.