D. 769546
L. 769553

UNIVERSITAT DE VALENCIA

FACULTAT DE PSICOLOGIA

EQUIVALENCIA PSICOMETRICA DE UNA TRADUCCION DEL CUESTIONARIO DE AUTOCONCEPTO FISICO PSDQ (PHYSICAL SELF-DESCRIPTION QUESTIONNAIRE) AL CASTELLANO

TESIS DOCTORAL

Presentada por:
INES TOMAS MARCO

Dirigida por:
VICENTE GONZALEZ ROMA
Profesor Titular de Metodología.

Valencia, Julio de 1998.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

A Gustavo

Las casas se construyen
piedra a piedra;
los sueños también.
Esta es la segunda piedra
de un sueño compartido,
que piedra a piedra,
se va haciendo realidad.

Quiero expresar mi agradecimiento más sincero a todas aquellas personas que me han ofrecido la colaboración y el apoyo necesario para poder llevar a cabo este trabajo.

En primer lugar mi agradecimiento a Vicente González Romá, no sólo por su orientación y ayuda en la realización de esta tesis, sino también por todo lo que me ha enseñado en el tiempo que llevamos trabajando juntos. Ha sido para mí un ejemplo de constancia, de optimismo, de tenacidad y autonomía en el trabajo. Trabajando con él he aprendido que los imposibles no existen.

También quiero expresar mi agradecimiento a mi familia, por estar siempre ahí, por su apoyo incondicional, su comprensión, y por su ayuda en todo lo que he necesitado.

Gracias a Ana, Begoña y Doris, por ser un punto de referencia y apoyo en los momentos bajos. Supongo que es cierto que "el que mejor comprende a un naúfrago es otro naúfrago". Un agradecimiento especial a Doris, compañera de fatigas en el proceso de recogida de la muestra, por compartir tantos kilómetros al volante, tantas jornadas inacabables de pase de cuestionarios, y alguna que otra anécdota divertida.

Tampoco puedo olvidarme de Lawrence Roche cuya colaboración, facilitando el acceso a los datos de una muestra australiana, ha sido fundamental en la realización de este trabajo. Gracias además por su paciencia en responder incansablemente desde las antípodas a todas las preguntas y cuestiones que le he planteado vía E-mail, y por propocionarme toda la información y publicaciones que he necesitado.

También quiero manifestar mi agradecimiento a todos los psicólogos, directores, jefes de estudios y profesores de los colegios e institutos en los que se ha realizado el pase de cuestionarios. Gracias por el interés que han demostrado en todo momento, y por la amabilidad y hospitalidad con la que se nos ha tratado.

A Patricia, Olga, Isabel, y a todos los amigos y amigas que con su apoyo constante y su comprensión me han hecho más llevadero tener que renunciar a muchos buenos momentos juntos.

Y muy especialmente a Gustavo, porque sin él, seguramente, este trabajo no tendría ningún sentido.

INDICE

CAPITULO 1

ESTUDIOS TRANSCULTURALES

La traducción y adaptación de instrumentos de medida cuenta con una larga tradición en el ámbito de la psicología. Uno de los primeros ejemplos data del año 1911. Nos referimos a la traducción que se hizo del francés al inglés de la Escala de Inteligencia para Niños Binet-Simon. En 1916, este test ya había sido traducido a 7 idiomas diferentes. Entre los instrumentos que han sido traducidos y adaptados a mayor número de idiomas podemos citar tests de aptitud intelectual como la Escala Wechsler de Inteligencia para Niños o la Escala de Inteligencia Stanford-Binet, y tests de personalidad como el Test de Apercepción Temática o el MMPI (Minnesota Multiphasic Personality Inventory) (Oakland y Hu, 1992).

¿Cuáles son las razones para traducir y adaptar instrumentos de medida a lenguas y culturas diferentes? Hambleton y Kanjee (1995), hablan de al menos tres buenas razones que han hecho que la traducción de cuestionarios sea hoy en día una práctica común:

1° Aumentar la fiabilidad y validez de los resultados obtenidos permitiendo a las personas ser evaluadas en la lengua de su elección.

2° Facilitar el desarrollo de estudios comparativos entre grupos culturales diferentes.

3° Reducir el coste económico y temporal que supondría el desarrollo de nuevos instrumentos.

En relación a la segunda de las razones que aportan estos autores, hay que señalar el creciente interés por el estudio y la comparación de culturas en los últimos 20 años. Cada vez más los investigadores tienden a buscar la confirmación de sus modelos y teorías más allá de sus propias fronteras geográficas. Pero antes de llegar a la comparación transcultural, han de tenerse en cuenta una serie de aspectos metodológicos que nos aseguren la validez de los resultados encontrados. Entre ellos podemos nombrar la necesidad de traducir y/o adaptar el instrumento de medida. La traducción del instrumento se realiza para resolver problemas de diferencia de idioma; la adaptación del instrumento se lleva a cabo fundamentalmente para evitar el sesgo cultural. Se puede afirmar por tanto, que la traducción/adaptación del instrumento de medida adquiere un papel relevante en los estudios de comparación transcultural.

El objetivo del presente trabajo es realizar la traducción y adaptación del cuestionario de autoconcepto físico PSDQ -"Physical Self Description Questionnaire"- (Marsh, Richards, Johnson, Roche y

Tremayne, 1994) a un idioma y cultura diferentes a aquellos en los que fue desarrollado en su versión original. Por tanto, dicho objetivo está enmarcado dentro de la problemática de la traducción/adaptación de instrumentos de medida. Como se ha comentado anteriormente, si se considera la traducción/adaptación de instrumentos como un requisito previo que permite el desarrollo de estudios de comparación entre culturas, la citada problemática se englobaría dentro de un marco más amplio: los estudios transculturales. Por ello, se ha considerado oportuno dedicar este primer capítulo a hacer una breve exposición sobre éstos. Qué son los estudios transculturales y cuál es su objetivo, cuál ha sido su evolución histórica y su relevancia actual, y los tipos de estudios transculturales, son algunos de los aspectos que van a ser tratados a continuación.

1. CONCEPTO, OBJETIVO, Y EVOLUCION HISTORICA DE LOS ESTUDIOS TRANSCULTURALES

Los estudios transculturales, tal y como su nombre indica, se refieren al estudio conjunto de diferentes culturas. En general, estos estudios pretenden analizar la existencia de similitudes o diferencias en las variables de interés, a través de las diferentes culturas analizadas. En el caso de confirmarse la existencia de diferencias, el siguiente paso sería interpretarlas, es decir, se deberían identificar aquellos aspectos de la cultura que son los responsables de dichas diferencias.

El origen de los estudios de comparación entre culturas lo encontramos en el desarrollo de la psicología transcultural. Esta disciplina cuenta con unos antecedentes históricos lejanos (Jahoda, 1992;

Cole, 1988, 1990), sin embargo, su surgimiento se localiza en un acercamiento interdisciplinar de cooperación conjunta entre la antropología y la psicología para estudiar la diversidad del desarrollo humano y de los procesos psicológicos (Shweder y Sullivan, 1993). A finales de los años 50, la psicología y la antropología no tenían prácticamente ningún tipo de relación. La psicología se dedicaba al trabajo experimental sobre aprendizaje animal, y la antropología al trabajo etnográfico (estudio de razas y pueblos) sobre rituales, mitos y relaciones de parentesco. La preocupación de la psicología se centraba en la búsqueda de leyes universales del funcionamiento psíquico, mientras que la antropología pretendía documentar las variaciones históricas y etnográficas de los diferentes colectivos sociales. Progresivamente, la antropología comienza a dirigir su interés hacia el individuo. Se desarrolla una etnografía centrada en la persona, y en el estudio de conceptos como self, mente, cuerpo, género, motivación o emociones (Briggs, 1970; Levy, 1973, 1978, 1983; Crapanzano, 1980; Heelas y Lock, 1981; Obeyesekere, 1981; Shostak, 1983; Weisner, 1984; Abu-Lughod, 1985, 1986). De forma simultánea, en el campo de la psicología comienza a reformularse el principio de la universalidad del psiquismo humano. Si se considera la cultura como "el conjunto de restricciones compartidas que limitan el repertorio conductual disponible de los miembros de cierto grupo" (Poortinga, 1992, pag. 10), se debe asumir que aunque los procesos psicológicos fueran universales, su expresión conductual va a estar limitada y modelada por la cultura en la cual se desarrollan. De esta forma, comienza a surgir una teoría del pluralismo psicológico. Se desarrollan seminarios, foros y actividades (ver Shweder, 1990; Shweder y Sullivan, 1990; 1993) que hacen que a finales de los años 80 el término "psicología cultural" sea cada vez más conocido y apunte al

surgimiento de una nueva subdisciplina dentro del campo de la psicología.

No debemos pensar sin embargo, que desde este enfoque se rechaza la existencia de universales en el funcionamiento psicológico. Como señalan Geertz (1984) y Kilbride (1992), la psicología transcultural debe ser al mismo tiempo "anti anti-relativista" y "anti anti-universalista". Los investigadores siguen alerta ante la posibilidad de generalizar los resultados encontrados en contextos particulares a contextos más amplios. Dicha generalización requiere evaluar los diferentes constructos, modelos y teorías psicológicas en culturas diferentes, buscando la replicación de los resultados. Sin embargo, el punto de partida teórico implica resaltar la influencia que la cultura tiene sobre el desarrollo de los procesos psicológicos y sobre el desarrollo humano, lo cual da origen a la diversidad. Podemos decir por tanto que la psicología transcultural pretende un estudio comparativo del modo en que la cultura y la psique se construyen mutuamente.

Los estudios pioneros de comparación entre culturas surgen fundamentalmente de la mano de investigadores que se encuentran bajo el paradigma de la psicología transcultural (Berry, 1967, 1969; Segall, Campbell, y Herskovits, 1966). Actualmente también son numerosos los estudios que desde la psicología transcultural pretenden estudiar diferentes constructos y procesos psicológicos: autoestima (Watkings y Dong, 1994), emociones (Mesquita y Frijda, 1992; Frijda y Mesquita, 1994), estereotipos de género (Williams y Best,1990; Best y Williams, 1994), atracción interpersonal (Shaver et al., 1991; Ellis et al., 1995), conflicto de rol, ambigüedad y sobrecarga (Peterson et al., 1995), etc. Una de las áreas de estudio donde se ha realizado mayor número de

comparaciones transculturales es la del rendimiento académico (Miura et al., 1993; Lapointe, Mead, y Phillips, 1989; Lapointe, Mead, y Askew, 1992). Estos estudios cobran una relevancia especial ya que ofrecen información valiosa para el desarrollo y perfeccionamiento de los programas educativos. Actualmente, la Asociación Internacional para la Evaluación del Rendimiento Educativo (IEA) (ver Keeves, 1992), está llevando a cabo el Tercer Estudio Internacional sobre Matemáticas y Ciencias (TIMSS). Dicho estudio comenzó en 1995 y se espera que sea concluido en 1999. Con la participación de más de 40 paises y la representación de 30 idiomas diferentes, es el estudio comparativo sobre rendimiento académico más amplio que se ha realizado hasta el momento.

Sin embargo, en la actualidad, gran parte de la investigación transcultural surge sin una planificación previa. Muchos investigadores comienzan realizando sus investigaciones en un determinado contexto cultural, y posteriormente consideran interesante ampliar su trabajo extendiendo la investigación a otras culturas diferentes (Van de Vijver y Hambleton, 1996). Estos investigadores no persiguen el desarrollo de un programa de investigación en psicología transcultural, simplemente realizan una ampliación de sus estudios. Por ejemplo, en la literatura encontramos múltiples ejemplos en los que el desarrollo de un instrumento que ha demostrado poseer una adecuada fiabilidad y validez en el contexto cultural en el que ha sido desarrollado y a partir del cual se han obtenido una serie de resultados interesantes, ha derivado en una adaptación posterior de este instrumento para ser aplicado en otras culturas diferentes y evaluar semejanzas o diferencias culturales (Hulin y Mayer, 1986; Smith, Tisak, Bauman, y Green, 1991; Ellis, Becker y Kimmel, 1993; Cheung, et al., 1996; Rubinstein, 1996). Se

puede afirmar por tanto, que el creciente interés en los últimos 20 años por los estudios de comparación entre culturas al que aludíamos anteriormente, se debe por una parte al propio desarrollo de la psicología transcultural, y por otra, a la tendencia de muchos investigadores de confirmar sus modelos y teorías en otras culturas. Esta diversidad de orientaciones será expuesta más ampliamente en el apartado siguiente cuando hablemos de los diferentes tipos de estudios transculturales.

Tal y como señalan Bond y Smith (1996), el crecimiento de la psicología transcultural se refleja en la reciente aparición de manuales sobre esta materia general (Berry et al., 1992; Brislin, 1993; Lonner y Malpass, 1994; Segall et al., 1990), y sobre otras materias más específicas, como la psicología social transcultural (Matsumoto, 1994; Moghaddam et al., 1993; Smith y Bond, 1994; Triandis, 1994), y la psicología organizacional transcultural (Adler, 1991; Erez y Earley, 1993; Triandis et al, 1993). También en la aparición de los volúmenes elaborados a partir de los congresos de la International Association for Cross-Cultural Psychology (Bouvy et al., 1994; Keats et al., 1989; Iwawaki et al., 1992; Pandey et al., 1995) y de una nueva edición del Handbook of Cross-Cultural Psychology (Berry et al., 1996). Por otro lado, el interés por los estudios comparativos también se pone de manifiesto con la profusión de publicaciones dedicadas exclusivamente a este tipo de estudios, como son el Journal of Cross-Cultural Psychology, el Hispanic Journal of Behavioral Sciences, o el Psychology and Developing Societies. Además, una revisión de la publicación electrónica PsycLit, también revela el incremento de los estudios de comparación transcultural (Van de Vijver & Lonner, 1995).

Pero, ¿qué repercusiones prácticas tiene el desarrollo de los estudios transculturales? Si reflexionamos por unos instantes sobre la realidad internacional que vivimos actualmente, no dudaremos en resaltar la relevancia de estos estudios. Observamos una tendencia a la apertura de fronteras y a la unificación de mercados, así como un creciente interés por el contacto y la cooperación entre naciones tanto en aspectos económicos, como educativos y culturales. Esta apertura internacional nos lleva a la necesidad de conocer más acerca de "los otros" y de establecer semejanzas y diferencias entre "los otros" y "nosotros mismos". Los estudios de comparación transcultural nos abren puertas para este conocimiento mutuo y suponen un enriquecimiento constante de los lazos de cooperación entre culturas diferentes. Además, tal y como señalan Shweder y Sullivan (1993), la realidad social actual requiere la consideración y el análisis de la diversidad cultural: "Quizá hace 30 o 40 años era razonable predecir que las tribus serían reemplazadas por los individuos, que las religiones serían reemplazadas por el conocimiento científico, y que la historia se orientaba en dirección hacia una cultura mundial homogénea de capitalistas consumidores que hablarían Esperanto (o Inglés). Hoy en día, éstas ya no son predicciones seguras (y quizá, ni siquiera razonables)" (p. 504).

2. TIPOS DE ESTUDIOS TRANSCULTURALES

Van de Vijver y Leung (1996) distinguen cuatro tipos de estudios transculturales en función de si la orientación es confirmatoria o exploratoria, y de si se consideran o no factores contextuales. Un estudio presenta una orientación confirmatoria cuando se parte de una hipótesis inicial sobre las diferencias entre los grupos culturales que se comparan;

el objetivo del estudio es confirmar dicha hipótesis. La orientación es exploratoria cuando no existe ninguna hipótesis inicial, y simplemente se indaga sobre la existencia o no de diferencias. En la siguiente tabla, tomada del trabajo de estos autores (pag. 288), se esquematiza la clasificación en base a las dos categorías citadas:

Consideración de factores contextuales	Orientación Confirmatoria	Orientación Exploratoria
NO	Estudios de Generalización	Estudios de Diferencias Psicológicas
SI	Estudios Guiados por una Teoría	Estudios de Validación Externa

Los estudios de generalización presentan una orientación confirmatoria, y no tienen en cuenta factores contextuales. Estos estudios parten de teorías, de instrumentos derivados de esas teorías, o de relaciones causales o correlacionales que han sido confirmadas en una determinada cultura. El objetivo del estudio es establecer la generalización de la teoría, del instrumento o de la relación. Por norma general, las investigaciones originales han sido realizadas en culturas occidentales, y pretenden ser generalizadas a otras culturas (occidentales o no), sin hacer referencia a factores culturales.

La mayor ventaja de los estudios de generalización es su versatilidad para poner a prueba la equivalencia de los resultados de culturas diferentes. Como se dispone de datos a priori con los que los nuevos datos van a ser comparados, es posible investigar varias hipótesis sobre las diferencias o las semejanzas transculturales. La debilidad de estos estudios radica en que no incluyen variables contextuales. Por lo tanto, si se encuentran diferencias entre las culturas

comparadas, a veces resulta complicado interpretarlas de forma adecuada.

La mayoría de los estudios de generalización que se encuentran en la literatura, son estudios en los que se aplica un instrumento derivado de una teoría. Algunos ejemplos son el trabajo de Schwartz (1992) realizado con datos procedentes de 20 paises, sobre la universalidad de la estructura de los valores humanos; los trabajos de Irvine (1979) y Vernon (1969, 1979) en los que se compara la estructura de la inteligencia en diferentes culturas; o los estudios que pretenden validar los cinco grandes factores de personalidad en diferentes culturas (McCrae y Costa, 1985; McCrae y John, 1992). Otro ejemplo de estudios de generalización es la replicación llevada a cabo por Amir y Sharon (1987) con estudiantes y universitarios israelíes de estudios socio-psicológicos realizados en culturas occidentales. El estudio de Leung (1987) sobre la elección de procesos de resolución de conflictos es un ejemplo de un estudio transcultural donde se pretende generalizar una relación causal.

Los estudios guiados por una teoría también adoptan un enfoque confirmatorio, pero a diferencia de los estudios de generalización, estos sí que tienen en cuenta los factores culturales. Se parte de la hipótesis de la existencia de diferencias culturales en la variable psicológica objeto de estudio, en base a que las culturas a comparar difieren en alguna dimensión relevante. Las diferencias culturales se establecen como hipótesis a priori que deben ser evaluadas. Estos estudios ponen a prueba una teoría sobre la relación entre variables culturales y variables psicológicas.

La ventaja principal de estos estudios es que informan directamente de la relación entre factores culturales y la variable psicológica objeto de estudio. Ello cobra más importancia si consideramos que este aspecto es generalmente considerado como el objetivo principal de la psicología transcultural (Berry, Poortinga, Segall, & Dasen, 1992). Su debilidad radica en que no consideran explicaciones alternativas a las postuladas inicialmente para dar cuenta de las diferencias transculturales encontradas.

Algunos ejemplos de este tipo de estudios son el trabajo de Berry (1976) y Berry et al. (1986), sobre los estilos cognitivos de los cazadores y recolectores de comida. En estos estudios, realizados con tribus del Africa Central, se pone de manifiesto que variables culturales, como los patrones educativos o la estructura social, influyen sobre los patrones de recolección de comida de un grupo cultural, originando diferencias entre los diferentes grupos comparados. También podemos citar el estudio de Earley (1989) sobre el "ganduleo social" (social loafing) con sujetos americanos y chinos. Este fenómeno hace referencia a que la gente trabaja menos cuando están en un grupo que cuando tiene que hacer la misma tarea individualmente. Partiendo de la diferencia entre las dos culturas en la variable individualismo-colectivismo, se confirmó que los sujetos americanos mostraban más ganduleo social que los chinos.

Los estudios de diferencias psicológicas adoptan una orientación exploratoria y no tienen en cuenta los factores contextuales. Son probablemente los más comunes en la literatura transcultural. Estos estudios implican la aplicación de un instrumento de medida en un nuevo contexto cultural; el propósito es explorar las diferencias transculturales, ya sea respecto a la puntuación promedio en la variable

psicológica medida, o a las propiedades psicométricas del instrumento. Generalmente, el instrumento original ha sido desarrollado en una cultura occidental, y su aplicación en otro grupo cultural pretende ser un paso hacia la difusión de su uso. No hay una teoría sobre la naturaleza de las diferencias transculturales, y tampoco se incluyen en el diseño los factores culturales, por lo que en el caso de encontrar diferencias, se buscan explicaciones a posteriori.

Esa es, según Van de Vijver y Leung, la ventaja de estos estudios: que no existe una hipótesis cerrada inicial sobre los factores contextuales que explicarían las diferencias transculturales en el caso de ser encontradas (como sí ocurría en los estudios guiados por una teoría). Esto permite una postura mucho más abierta a la hora de buscar explicaciones sobre la presencia de diferencias transculturales. La debilidad de estos estudios radica en que, como no se tienen en cuenta variables contextuales, muchas veces la interpretación de diferencias transculturales a posteriori, puede resultar complicado. Además puede ocurrir el "fenómeno de pesca" ("fishing") descrito por Cook y Campbell (1979). Este fenómeno ocurre cuando se aplican un gran número de pruebas estadísticas para evaluar la hipótesis nula de la no existencia de diferencias entre grupos culturales. Estos procedimientos múltiples de evaluación (o fenómeno de pesca de la significación estadística) pueden llevar fácilmente a rechazar la hipótesis nula, siendo ésta verdadera, y por lo tanto, conducir a conclusiones incorrectas sobre la ocurrencia de diferencias transculturales.

La mayoría de los artículos que aparecen publicados en el Journal of Cross-Cultural Psychology son de este tipo. Un ejemplo es el trabajo

realizado por Guida y Ludlow (1989) comparando la ansiedad ante los exámenes de escolares americanos y chilenos.

Por último, los estudios de validación externa pretenden explorar el significado y las causas de diferencias transculturales con la ayuda de factores contextuales. En este tipo de estudios no suele haber hipótesis específicas a priori, sino que parten de la inclusión de un gran número de variables contextuales de forma exploratoria. Si se encuentran diferencias entre los grupos culturales, se intentará identificar una interpretación adecuada de esas diferencias en base a las variables contextuales que han sido medidas. Dichas variables pueden proceder de diferentes niveles de análisis: individual (por ejemplo, la puntuación de los sujetos en un test de individualismo-colectivismo), intermedio (por ejemplo, datos de la familia o de la escuela), o nivel cultural (como por ejemplo, la densidad de población, el producto nacional bruto, o la renta nacional). En algunos casos, la validación externa se basa en estudios previos de generalización o de diferencias psicológicas en los que se informa de diferencias transculturales, y en otros casos, la observación de diferencias transculturales y la validación externa se combinan en el mismo estudio.

La ventaja de estos estudios es que se centran directamente en la interpretación de las diferencias transculturales, es decir, en la búsqueda de explicaciones para las diferencias encontradas. Esto es importante, ya que este aspecto suele descuidarse a menudo en la psicología transcultural, y muchas veces se informa de la existencia de diferencias, pero no se profundiza en la búsqueda de explicaciones. Su punto débil es que la elección de las variables que den cuenta de estas diferencias es arbitraria, por lo que éstas pueden ser irrelevantes desde

un punto de vista psicológico. Por ejemplo, se ha visto que la distancia de la capital de un país al ecuador es un buen predictor de la puntuación en varios tests psicológicos (Van de Vijver y Leung, 1996). Sin embargo, es obvio que estos resultados no ofrecen mucha información sobre las variables psicológicas subyacentes a las diferencias en la ejecución.

Podemos señalar como ejemplo de esta aproximación el trabajo de Bond (1991) sobre valores y salud, y el trabajo de Williams y Best (1982) realizado con datos de 30 paises sobre estereotipos de género. Estos autores comienzan demostrando la existencia de diferencias transculturales en el constructo medido, y posteriormente interpretan estas diferencias en base a una serie de variables de nivel cultural, como el producto nacional bruto, o el gasto per capita en educación y salud.

Por último, cabe señalar que el objetivo perseguido en cada uno de estos estudios va a determinar la necesidad de utilizar diferentes metodologías y procedimientos de análisis de datos.

3. EL PROBLEMA DEL SESGO EN LOS ESTUDIOS TRANSCULTURALES

En los estudios de comparación entre culturas se utilizan instrumentos de medida para la recogida de datos del constructo a evaluar. Se puede afirmar por tanto, que el instrumento de medida adquiere un papel relevante en los estudios transculturales. Sin embargo, la aplicación de un instrumento de medida en una cultura diferente, en la que se utiliza una lengua diferente, no se reduce a la simple producción de un texto en otro idioma, su administración, y la comparación de resultados. Es importante tener en cuenta la posible

amenaza de tres tipos de sesgo que deben ser evitados para que no invaliden los resultados de la investigación. Los tres tipos de sesgo que aparecen documentados en la literatura transcultural son: el sesgo del constructo, el sesgo del método y el sesgo de los items (Van de Vijver y Hambleton, 1996; Van de Vijver y Leung, 1996).

El sesgo del constructo hace referencia a la no equivalencia del constructo en los diferentes grupos culturales. Cuando el constructo psicológico objeto de estudio no es idéntico en las culturas comparadas, se corre el riesgo de que este sesgo invalide los resultados del estudio. Es importante tener en cuenta que si el instrumento de medida ha sido desarrollado en uno de los grupos, probablemente será adecuado para medir el constructo en ese grupo cultural. Sin embargo esto no garantiza que vaya a representar de forma adecuada el constructo en los otros grupos culturales.

Por su parte, el sesgo del método se refiere a problemas en la elaboración o administración del instrumento. Cuando este tipo de sesgo ocurre, el instrumento representa de forma adecuada el constructo psicológico a medir, pero es el procedimiento de medida el que introduce diferencias entre los grupos debido a que éstos no tienen el mismo grado de familiarización con los materiales utilizados, con el formato de los items, etc.

Finalmente, el sesgo de los items, hace referencia a la no equivalencia psicométrica de los items, y por lo tanto, a la no equivalencia entre las diferentes versiones del instrumento de medida. Normalmente este sesgo es debido a una traducción inadecuada o a una elección incorrecta de las palabras.

Para poder interpretar los resultados de una investigación transcultural de forma adecuada, es necesario que las diferencias o semejanzas encontradas sean debidas a diferencias o semejanzas reales, y no sean efecto de algún tipo de sesgo. La pregunta a responder sería: ¿cómo podemos evitar el efecto de estos tres tipos de sesgo y asegurarnos de que las versiones del instrumento utilizadas en cada cultura son equivalentes? La respuesta es doble, y podría dar lugar a una clasificación de los métodos en preventivos y confirmatorios. Los métodos preventivos serían todos aquellos que bajo el título general de "adaptación del instrumento de medida" pretenden prevenir el efecto del sesgo, y asegurar la equivalencia de los instrumentos. Por otro lado, los métodos confirmatorios serían todos aquellos, que una vez recogidos los datos y antes de pasar a analizar las semejanzas o diferencias transculturales, se aplican para garantizar que la adaptación ha sido adecuada, y por lo tanto, que los resultados obtenidos son válidos.

Vamos a dedicar los dos capítulos siguientes a la exposición de los diferentes métodos que aparecen en la literatura transcultural, y que permiten prevenir o estudiar la existencia de sesgo en los estudios de comparación entre culturas. En el capítulo 2 hablaremos de los métodos preventivos: presentaremos los procedimientos para llevar a cabo una adaptación adecuada, así como las recomendaciones necesarias para evitar que el proceso de adaptación del instrumento se vea afectado por alguno de los sesgos comentados anteriormente; también hablaremos de una serie de aspectos a tener en cuenta en la aplicación del instrumento que ayudan a reducir el riesgo de sesgo. En el capítulo 3 expondremos los métodos utilizados para estudiar la presencia de sesgo en los estudios de comparación transcultural, haciendo especial hincapié en aquellos

que permiten evaluar la equivalencia de las diferentes versiones del instrumento.

CAPITULO 2

METODOS PARA EVITAR EL SESGO

Para poder comparar culturas diferentes respecto a una variable psicológica determinada se necesita, en primer lugar, un instrumento que mida dicha variable, y en segundo lugar, contar con versiones equivalentes del instrumento que estén adaptadas para recoger información válida y fiable en los diferentes contextos culturales a comparar. Por lo tanto, a la hora de adaptar un instrumento de medida, deben seguirse todos los procedimientos metodológicos que aseguren la correcta adaptación del instrumento a las características de los diferentes grupos culturales.

En 1993, la Comisión Internacional de Tests (International Tests Comission, ITC) elaboró 22 directrices prácticas para la adaptación de tests. Estas directrices cubren cuatro dominios: el contexto, la construcción y adaptación del test, la aplicación, y la documentación e interpretación de resultados. Las directrices referidas al contexto, recogen dos principios básicos de los estudios transculturales que hacen referencia a la necesidad de minimizar los efectos de las diferencias culturales que no sean relevantes para los objetivos del estudio, y a la necesidad de evaluar la equivalencia de los constructos en los diferentes grupos culturales. Respecto a la construcción y adaptación del test, encontramos diez recomendaciones para el desarrollo de instrumentos que se utilizan en investigación transcultural, incluyendo aspectos como la consideración de las diferencias lingüísticas, el uso adecuado del vocabulario, el problema de la familiaridad con los materiales y procedimientos, o la necesidad de emplear diseños y técnicas estadísticas adecuados para poder evaluar la equivalencia de las diferentes versiones del instrumento. Las seis directrices dedicadas a la administración de los instrumentos hacen referencia a aspectos como la necesidad de asegurar la similitud de las condiciones de aplicación, o aspectos relativos a la selección de los aplicadores. Finalmente, las cuatro directrices relativas a la documentación e interpretación de resultados recogen aspectos relacionados con la interpretación y la comparación de puntuaciones de diferentes grupos culturales.

Estas directrices representan un punto de referencia necesario para cualquier investigador que pretenda realizar un estudio de comparación entre culturas. A continuación vamos a desarrollar las directrices relativas a la construcción y adaptación de tests, y las relativas a la aplicación. Sin embargo, para dar mayor agilidad y

coherencia a la lectura de este capítulo, hemos preferido no hacerlo de forma sistemática a modo de listado como aparecen en la fuente original, sino incluir estas directrices dentro del esquema general de nuestra exposición, intercalándolas con las recomendaciones y aportaciones de otros autores respecto a los procedimientos a seguir en el proceso de adaptación y aplicación del instrumento de medida. Remitimos al lector interesado en una exposición sistemática, al trabajo de Hambleton (1994); o para un análisis más detallado, a los trabajos de Van de Vijver y Hambleton (1996), y Hambleton (1996), donde se recoge el fundamento y la explicación de la inclusión en el informe de la ITC de cada una de estas 22 directrices.

1. PROCESO DE CONSTRUCCION Y ADAPTACION DE INSTRUMENTOS DE MEDIDA

Antes de pasar a desarrollar el contenido de este apartado, vamos a hacer una breve aclaración terminológica. En la literatura anglosajona, de forma tradicional, se han utilizado indistintamente los términos traducción (translating) y adaptación (adapting), para referirse al proceso general de desarrollo de un instrumento que pueda ser aplicado en un contexto cultural determinado a partir de otro instrumento que ha sido creado en otro contexto cultural diferente. Prueba del uso no diferenciado de estos dos términos, son los trabajos de Hambleton en los que las mismas directrices, comentadas anteriormente, figuran bajo encabezados diferentes: "Guidelines for adapting educational and psychological tests" (Hambleton, 1994), y "Guidelines for translating tests" (Van de Vijver y Hambleton, 1996). Desde nuestro punto de vista, sería conveniente delimitar la terminología, ya que el concepto de

traducción es más restringido, e indicaría solamente un aspecto del proceso general de adaptación en el caso de que los grupos culturales no utilicen el mismo idioma. Esta misma idea la encontramos en un trabajo posterior de Hambleton, donde el mismo autor señala: "... se ha preferido utilizar el término adaptación en vez del más popular de traducción, debido a que es más amplio y más representativo de lo que ocurre en la práctica cuando se prepara un test para utilizarlo en un segundo idioma y/o cultura" (Hambleton, 1996, pag. 211). Sin embargo, esta delimitación terminológica, todavía no está exenta de problemas. Como veremos más adelante, Van de Vijver y Leung (1996) y Van de Vijver y Hambleton (1996), nos hablan de tres procedimientos diferentes a emplear cuando trabajamos con un test y queremos utilizarlo en otra cultura: aplicar, adaptar, o construir un nuevo instrumento. En este contexto, la adaptación hace referencia a que el cuestionario necesita sufrir alguna modificación respecto al cuestionario original para poder ser aplicado al grupo cultural objetivo. Los otros dos procedimientos indicarían, en el caso de la aplicación, que no es necesario realizar ninguna modificación, o por el contrario, en el caso de la construcción de un nuevo instrumento, que sería necesario elaborar uno totalmente nuevo. Aquí utilizaremos el término adaptación para referirnos al proceso general que implica la aplicación de un instrumento en un grupo cultural diferente, y que incluye entre otros aspectos, la traducción del instrumento en caso de ser necesario, y la decisión sobre qué procedimiento es más adecuado utilizar (aplicar, adaptar, o construir un nuevo instrumento).

1.1. Procedimientos para la adaptación de instrumentos de medida

En los estudios transculturales es frecuente encontrar instrumentos de medida que han sido elaborados en un determinado contexto cultural, y que posteriormente van a ser utilizados en otra cultura diferente para recoger información que permita la comparación entre culturas. En estos casos se debe tener en cuenta que una aplicación directa del instrumento no siempre es adecuada. Se debería evaluar previamente la idoneidad del instrumento de medida para ser aplicado en el nuevo contexto cultural, y realizar modificaciones en caso de que sea necesario. Dependiendo de los cambios requeridos, se encuentran en la literatura transcultural tres procedimientos posibles para la adaptación del instrumento de medida: aplicar, adaptar o construir un instrumento nuevo (Van de Vijver y Hambleton, 1996; Van de Vijver y Leung, 1996). Como veremos a continuación, la decisión sobre la elección de uno de estos tres procedimientos a la hora de realizar la adaptación del instrumento, se encuentra estrechamente relacionada con los tres tipos de sesgo que hemos comentado en páginas anteriores (del constructo, del método y de los items).

1.1.1. Aplicación

Este procedimiento implica que no se realiza ninguna modificación sobre el instrumento de medida. En el nuevo contexto cultural se aplica el instrumento original, o una versión traducida en caso de que las dos culturas no compartan el mismo idioma. La similitud entre culturas hace que se asuma la ausencia de sesgos del constructo y del método. Unicamente se examina el sesgo de los items. Esta es la opción más

sencilla, quizá por ello sea también la más utilizada, pero únicamente se puede utilizar cuando las dos culturas a comparar son muy similares. Ejemplos de su uso los encontramos en los trabajos de Brislin (1980), y Werner y Campbell (1970).

1.1.2. Adaptación

Este procedimiento se utiliza para evitar el sesgo del método. En aquellos casos en los que algún aspecto de forma del cuestionario original, como el formato de respuesta, los tipos de estímulos presentados, etc., no son familiares en el nuevo contexto cultural en el que va a ser aplicado el instrumento, es necesario realizar alguna modificación que evite ese sesgo desfavorable. También cuando el constructo no está totalmente cubierto en el nuevo grupo, se puede hacer una adaptación tomando partes del instrumento original, y modificando, reemplazando o añadiendo items que midan aquellos aspectos que no se contemplaban en el instrumento original. Sin embargo, hay que tener en cuenta que la administración de un test que no es totalmente idéntico complica los análisis estadísticos.

Un ejemplo de este procedimiento lo encontramos en el trabajo de Van Haaften y Van de Vijver (in press), en el que se aplica el inventario Coping Strategy Indicator de Amirkhan (1990). El ítem "ver la televisión más de lo usual", tuvo que ser eliminado porque en el área en la que el instrumento iba a ser aplicado no había electricidad, por lo que tampoco había televisores.

1.1.3. Construcción de un instrumento completamente nuevo

Este procedimiento se utiliza cuando se considera que el instrumento original es totalmente inadecuado para ser administrado en el nuevo contexto cultural. Es decir, se utiliza para evitar el sesgo del constructo. En estos casos, el investigador está más interesado en analizar cuáles son las características del constructo medido en cada grupo, y no tanto en estudiar las diferencias o similitudes.

A la hora de decidir cuál de los tres procedimientos emplear, se deben considerar las implicaciones teóricas y prácticas que cada uno de ellos presenta. Van de Vijver y Leung (1996) proponen adoptar por defecto la opción de aplicar el mismo instrumento. Argumentan su propuesta señalando las siguientes ventajas:

1ª La aplicación del mismo instrumento, sin realizar ningún tipo de modificación posibilita comparar los resultados obtenidos en una investigación con otros resultados ya publicados en la literatura.

2ª La aplicación del mismo instrumento también posibilita mantener la equivalencia escalar, equivalencia que no tendríamos si comparamos los resultados del instrumento original con los resultados obtenidos con un instrumento completamente nuevo.

3ª Administrar un instrumento ya existente supone un ahorro económico y de esfuerzo en comparación con la opción de desarrollar y evaluar las propiedades psicométricas de un instrumento adaptado o nuevo.

Sin embargo, se debe tener en cuenta que aplicar un instrumento ya existente no es siempre la mejor elección. Si el instrumento no cubre aspectos importantes del constructo psicológico bajo estudio, o si muestra claros sesgos etnocéntricos, la adaptación o la construcción de un nuevo instrumento serían mejores opciones. La decisión debe ser el resultado de un análisis de costos-beneficios, con el tiempo y el dinero como costos y la representación del constructo y la evitación de sesgos como beneficios.

Evitar el sesgo del constructo es uno de los criterios que debería de guiar la decisión a la hora de decidir qué alternativa elegir. Si el constructo objeto de estudio no es idéntico en los diferentes grupos culturales, quedará descartada la opción de aplicar el mismo instrumento. Para evitar el sesgo del constructo, será necesario realizar las modificaciones y adaptaciones necesarias para que el instrumento mida de forma adecuada el constructo en los nuevos grupos culturales; o crear un nuevo instrumento en el caso de que las diferencias así lo requieran. La cuestión a plantearse es si el instrumento recoge de forma adecuada y suficiente el constructo psicológico a medir. Para responder a este interrogante, se requiere un conocimiento adecuado de los contextos culturales en los que el instrumento va a ser aplicado. Una forma de evitar el sesgo del constructo es realizar estudios piloto con la cooperación de expertos miembros de la cultura objetivo, para identificar las características conductuales del constructo en este nuevo contexto cultural.

1.2. Criterios a tener en cuenta en la construcción de instrumentos de medida

Por norma general, los instrumentos de medida que se utilizan en los estudios transculturales, han surgido en un contexto cultural determinado, sin que el autor o autores tuvieran ninguna intencionalidad previa de generalizar el uso del cuestionario en otras culturas diferentes. Posteriormente se ha planteado su adaptación a otro contexto diferente para realizar estudios comparativos. En este caso se plantea una adaptación a posteriori, es decir, tras la construcción y utilización del instrumento en la cultura original. El planteamiento cambia si desde el principio, la construcción del instrumento de medida se realiza teniendo en cuenta que va a ser utilizado en diferentes culturas. En este caso, el proceso de adaptación podría comenzar desde el principio de la construcción del cuestionario. ¿Qué criterios deberían de tenerse en cuenta a la hora de elaborar un instrumento de medida del que se sabe de antemano que va a ser utilizado en diferentes contextos culturales? Hambleton (1996) y Van de Vijver y Hambleton (1996) proponen una serie de aspectos importantes a considerar en el proceso de construcción de un instrumento que va a ser utilizado en investigación transcultural. La consideración de estas recomendaciones ayuda a reducir el riesgo de sesgos del método y de los items.

1.2.1. Elección del formato de los items

La falta de familiaridad de uno de los grupos con el formato de los items puede significar una clara desventaja y ofrecer por tanto medidas sesgadas en ese grupo (este sería un ejemplo de sesgo del método). Por tanto, se debería elegir un formato que sea familiar a los grupos que se

van a comparar, y utilizar solamente aquellos formatos en los que todos los grupos tienen experiencia. En el caso de que cada grupo tenga experiencia en formatos diferentes, se puede buscar un equilibrio combinándolos. Por ejemplo, si uno de los grupos culturales tiene experiencia en items de elección múltiple, y el otro en items de respuesta corta, se podría elaborar un instrumento compuesto por items de los dos formatos en la misma proporción. Los formatos preferibles (en el caso de que sean familiares para todos los grupos) son los items de elección múltiple, o las escalas sencillas de calificación como las de tipo Likert.

1.2.2. Elección de los materiales

También es importante considerar la familiaridad con los materiales utilizados para la administración del instrumento, con el fin de evitar el sesgo del método. Es importante cuidar que los procedimientos de evaluación (por ejemplo, lápiz y papel, ordenador), los estímulos (como por ejemplo, diagramas, tablas, figuras, etc.), e incluso el tipo de instrucciones utilizadas, sean familiares a los diferentes grupos culturales en los que se va a aplicar el instrumento.

1.2.3. Precauciones lingüísticas

Se debería elegir el vocabulario y la estructura de las frases de tal modo que faciliten su traducción posterior. Como comentábamos anteriormente, los problemas de traducción suelen originar falta de equivalencia entre los items, es decir, sesgo de los items. Se recomienda también evitar las unidades de medida y de temperatura, dado que varían de un país a otro.

1.2.4. "Descentramiento" (decentering)

Cuando las dos culturas objeto de estudio no tienen el mismo idioma, es necesario realizar una traducción. Werner y Campbell (1970) proponen el descentramiento como una medida para ajustar las versiones original y traducida del instrumento. Esta estrategia es posible cuando el instrumento se está desarrollando en el idioma fuente (versión original), y a la vez se está llevando a cabo la versión en el idioma objetivo (versión traducida). De este modo, cuando ciertas palabras o expresiones no tienen equivalente en el idioma objetivo, se hacen las revisiones y modificaciones oportunas del instrumento en el idioma fuente para utilizar versiones equivalentes en los dos idiomas.

1.3. Traducción de instrumentos de medida

Como ya comentamos en el capítulo anterior, Hambleton y Kanjee (1995) resaltan la importancia de la traducción de instrumentos de medida en base a tres argumentos: 1° aumentar la fiabilidad y validez de los resultados obtenidos permitiendo a las personas ser evaluadas en la lengua de su elección, 2° facilitar el desarrollo de estudios comparativos entre grupos culturales diferentes, y 3° reducir el coste económico y temporal que supondría el desarrollo de nuevos instrumentos. El primer argumento cobra gran importancia actualmente debido al creciente reconocimiento del bilingüismo. Evaluar a un sujeto en su segunda o tercera lengua puede originar sesgos que atenten contra la fiabilidad y la validez de los resultados. Este hecho repercute principalmente en estudios realizados en paises o regiones donde existen al menos dos lenguas oficiales.

En este trabajo vamos a centrarnos en el segundo argumento. En un estudio transcultural, cuando las culturas a comparar tienen idiomas diferentes, resulta necesario traducir el instrumento de medida. La traducción representa en estos casos un eslabón fundamental para asegurar la obtención de medidas equivalentes que puedan ser comparadas. Esta equivalencia puede verse en peligro por múltiples causas: por ejemplo, si se realiza una traducción que no se ajusta al significado del instrumento en la lengua original, o si la traducción es tan literal que resulta difícil de entender en el idioma objetivo. Todo ello se traduce en lo que Van de Vijver y Hambleton (1996) denominan sesgo de los items. Es decir, que los items de las dos versiones del instrumento no sean equivalentes, y por lo tanto estén midiendo cosas diferentes. Para evitar todos estos problemas, en el proceso de traducción se deben de tener en cuenta aspectos tan importantes como: quién va a realizar la traducción, cómo se va a llevar a cabo dicha traducción, el nivel de familiaridad de los individuos de los dos grupos culturales con el vocabulario utilizado, etc. Veamos a continuación todos estos aspectos.

1.3.1. Elección de los traductores

Según diferentes autores (Hambleton, 1996; Hambleton y Kanjee, 1995), es importante contar con un equipo de traductores, ya que una traducción realizada por varias personas siempre será más rica y completa que la realizada por una sola. El trabajo en equipo permite contemplar diferentes puntos de vista y ofrece un mayor número de posibilidades para resolver los problemas que puedan surgir durante la traducción.

Por otro lado, es conveniente que los traductores sean competentes en los dos idiomas y que conozcan muy bien las dos culturas, especialmente la cultura objetivo, es decir, la cultura del idioma al cual se está adaptando el instrumento.

También es importante que los traductores conozcan la materia sobre la que trata el instrumento, y que tengan unos conocimientos básicos sobre la construcción de tests y la elaboración de items. De no ser así, se les debería proporcionar unas sesiones previas de formación para la adquisición de estos conocimientos, y que garantizen por tanto que van a tener la preparación adecuada.

1.3.2. Consejos para optimizar la traducción

Brislin, Lonner y Thorndike (1973), y posteriormente Brislin (1980; 1986) proponen una serie de consejos para asegurar la adecuada traducción del instrumento de medida. Estos son aplicables tanto en la construcción del instrumento original para favorecer su posterior traducción, como en el propio proceso de traducción. Los principales consejos que proponen estos autores figuran a continuación.

* Usar sentencias cortas y evitar las palabras innecesarias.
* Usar la voz activa en lugar de la pasiva.
* Repetir los nombres en lugar de usar pronombres ya que estos últimos pueden ser más vagos.
* Evitar metáforas o coloquialismos, que pueden crear problemas de traducción.
* Evitar verbos y preposiciones de tiempo o de lugar que sean vagos o imprecisos (por ejemplo: pronto, a menudo).

* Evitar las formas en posesivo cuando sea posible ya que puede ser difícil determinar los propietarios.
* Usar términos específicos y no generales (por ejemplo, el término " los miembros de tu familia" puede diferir enormemente entre culturas).
* Evitar el modo subjuntivo en la conjugación de los verbos.
* Cuando se introduzca algún concepto clave o información importante, añadir frases que lo enfatizen y resalten su importancia.
* Evitar frases con verbos que sugieran acciones diferentes.

En la literatura transcultural encontramos todavía otros aspectos a tener en cuenta en el proceso de traducción del test para reducir el riesgo de incurrir en el sesgo de los items. Por ejemplo, el conocimiento de la frecuencia de uso de las palabras en cada uno de los idiomas puede ser de gran valor para obtener adaptaciones válidas. También es importante cuidar que el vocabulario utilizado en las diferentes versiones del instrumento sea comparable en cuanto al nivel de dificultad de las palabras, legibilidad, gramática, y estilo de escritura.

2. APLICACION DEL INSTRUMENTO ADAPTADO

En este apartado incluimos los consejos y recomendaciones que deberían tenerse en cuenta en la aplicación del instrumento adaptado en el nuevo grupo cultural de interés para reducir el riesgo de sesgo, haciendo referencia en este caso al sesgo de método principalmente. Según Van de Vijver y Poortinga (1991, 1992) los principales aspectos a considerar en la aplicación del instrumento adaptado son los siguientes:

1) características personales de los aplicadores;

2) interacción entre el aplicador y los sujetos de la muestra;

3) procedimientos de respuesta;

4) familiaridad con los estímulos.

Vamos a ver cada uno de ellos con más detalle a continuación.

2.1. Características personales de los aplicadores

Parece evidente que la persona o personas encargadas de la aplicación del instrumento pueden influir sobre la validez de los resultados, sobre todo si pertenecen a una cultura diferente a la de los sujetos de la muestra. Este fenómeno ha sido estudiado sistemáticamente y demostrado por diferentes autores (Jensen, 1980; Super, 1981). Además, ciertas características del aplicador, como el género, la edad, o incluso la forma de vestir, pueden influir en el resultado de la medición. Para evitar este tipo de influencia, Hambleton (1996) propone que los aplicadores sean elegidos de la población a la que se va a aplicar el instrumento. En el caso de que esto no sea posible, con la ayuda de informadores locales se deberían precisar las características de los aplicadores que pueden poner en peligro la validez de los resultados. Al mismo tiempo, se debe procurar que los aplicadores estén familiarizados con la cultura y con el idioma de los sujetos de la muestra. Este autor también recomienda que los aplicadores tengan experiencia en la aplicación de tests; de no ser así, se les debería proporcionar un entrenamiento básico.

2.2. Interacción entre el aplicador y los sujetos de la muestra

La aplicación del instrumento supone una interacción entre el aplicador o aplicadores y los sujetos de la muestra. Cuando no compartan el mismo idioma, es posible que surjan problemas de comunicación. En estos casos, Van de Vijver y Leung (1996) proponen, si es posible, reducir al mínimo la comunicación verbal; y si se decide administrar el instrumento con la ayuda de algún intérprete, aconsejan también evaluar la influencia potencial de éste. Los consejos de Van de Vijver y Poortinga (1991) para evitar que esta interacción afecte a la validez de la investigación, son que el aplicador evite influir sobre los examinados siguiendo al pie de la letra las instrucciones. Previamente debe asegurarse de que éstas estén correctamente adaptadas y sean claras y exhaustivas, dejando un margen mínimo para las explicaciones verbales y así evitar problemas de comunicación entre aplicador y examinados.

2.3. Procedimientos de respuesta

Otro aspecto importante a tener en cuenta en la aplicación del instrumento, son los procedimientos de respuesta. Estos ya han sido comentados anteriormente al hablar del proceso de construcción del test en la lengua original. Sin embargo, dada su importancia, y ya que no siempre se trabaja con instrumentos que hayan sido elaborados teniendo en cuenta estos consejos, volvemos a incluirlos en este apartado como un aspecto a considerar en la aplicación del instrumento.

Respecto a los procedimientos de respuesta, debemos comprobar que tanto el formato de respuesta de los items (por ejemplo, formato tipo Likert), como los materiales utilizados (por ejemplo, lápiz y papel,

ordenador), son familiares para los grupos a evaluar, y que éstos poseen una experiencia similar a la del grupo de origen. Si no existe esta familiarización, es importante dedicar un tiempo antes del pase del instrumento para familiarizar a los sujetos con estos procedimientos (Hambleton, 1996; Van de Vijver y Leung, 1996).

2.4. Familiaridad con los estímulos

Por último, otro aspecto que no debemos olvidar en la aplicación del test, es la familiaridad de los sujetos de la muestra con los estímulos. Esta es la fuente de sesgo en comparaciones entre grupos culturales diferentes que más se ha estudiado, y la que se ha mencionado más frecuentemente como causa de sesgo (Irvine y Carroll, 1980). Este aspecto es especialmente importante en determinado tipo de instrumentos, como por ejemplo los tests cognitivos, en los que frecuentemente se presenta a los sujetos diferentes estímulos (objetos, fotografías, figuras) que deben manipular (ordenar, emparejar, etc.). De ahí que en los test cognitivos se tienda cada vez más a utilizar figuras geométricas simples (cuadrados, triángulos, círculos) como estímulos, ya que parece asumirse una familiarización compartida entre diferentes culturas con este tipo de estímulos. En otros instrumentos, en los que los estímulos son los items que el sujeto debe contestar, el problema de la familiarización se debería a aspectos ya mencionados anteriormente, como el uso de vocabulario complejo o de expresiones poco usuales, que originarían problemas de comprensión.

Hambleton (1996) señala otros aspectos a tener en cuenta para asegurar la correcta aplicación del instrumento. Este autor aconseja evitar el establecimiento de tiempo límite; no todos los grupos culturales

tienen la misma experiencia en tests de velocidad (Van Leest y Bleichrodt, 1990), por lo que es conveniente evitar que la velocidad sea un factor influyente en la ejecución. Las condiciones de aplicación del instrumento también pueden ser una fuente de variación indeseada de las puntuaciones; se debería intentar que éstas sean lo más parecidas posibles en todos los grupos culturales en los que el instrumento sea aplicado. Además de todas las comentadas anteriormente, otra fuente de sesgo puede ser la existencia de diferencias entre las muestra fuente y objetivo en aspectos tales como la deseabilidad social, el patrón de respuesta (por ejemplo, aquiescencia), o la motivación de los sujetos para contestar la prueba.

En este capítulo hemos hablado de los diferentes procedimientos y recomendaciones que deben considerarse durante el proceso de construcción y adaptación de instrumentos de medida, y durante su posterior aplicación, para evitar el efecto del sesgo en los estudios de comparación entre culturas. Las consideración de éstas recomendaciones permite prevenir el efecto de los sesgos que amenazan la validez de la investigación (sesgo del constructo, del método y de los items). Sin embargo, su uso y consideración no exime de la necesidad de evaluar la presencia de sesgo. Existen diferentes métodos para el estudio del sesgo en la investigación transcultural. Su uso permite comprobar si los métodos preventivos han sido efectivos para evitar el efecto de los sesgos en los resultados de la investigación, o si por el contrario, y a pesar de las precauciones tomadas, el estudio se ve afectado por algún tipo de sesgo. En el capítulo 3 presentamos una exposición detallada de los diferentes procedimientos que aparecen en la literatura transcultural para evaluar la existencia de sesgo, haciendo especial hincapié en

aquéllos que permiten analizar la equivalencia entre las diferentes versiones del instrumento.

CAPITULO 3

METODOS PARA EL ESTUDIO DEL SESGO EN LA INVESTIGACION TRANSCULTURAL

El problema del estudio del sesgo se remonta a principios del siglo XX con los trabajos de Alfred Binet sobre el cociente intelectual (Binet y Simon, 1916/1973). Binet detectó al aplicar pruebas de cociente intelectual a niños de bajo nivel socioeconómico que éstos obtenían por norma general resultados más bajos. Un estudio más profundo le llevó a concluir que estas diferencias se debían a la presencia de items que medían el efecto del entrenamiento cultural que se producía en casa o en la escuela, más que la capacidad mental. Esto le llevó a eliminar ciertas categorías de items que podían estar perjudicando a los grupos de menor nivel socioeconómico, ya que por problemas de escolarización o debido al tipo de interacción familiar, podían tener unas experiencias de entrenamiento cultural más pobres.

Sin embargo, el estudio del sesgo en los instrumentos de medida cobra especial interés en la literatura psicométrica, educativa y de evaluación a finales de los años 60. Es entonces cuando en Estados Unidos aparecen diferentes movimientos de reivindicación de la igualdad de derechos. Estos movimientos denunciaban la necesidad de utilizar instrumentos psicométricos que no perjudicaran a los grupos culturales, raciales o étnicos minoritarios. Es necesario señalar que ya entonces los tests psicométricos jugaban un papel fundamental en la admisión a instituciones educativas o a puestos de trabajo. Los instrumentos que se utilizaban en estas situaciones de selección educativa y laboral eran mayoritariamente pruebas de cociente intelectual. En este tipo de pruebas se había observado que las poblaciones de sujetos negros e hispanos presentaban un rendimiento considerablemente inferior a la población de sujetos blancos. La explicación parecía ser, como ya habían puesto de manifiesto los trabajos de Binet, que estos instrumentos contenían conocimientos y habilidades que formaban parte de la cultura de la población blanca de clase media, mientras que los grupos minoritarios (blancos e hispanos) tenían menos ocasiones y oportunidades de aprender estos contenidos. Además, el idioma jugaba también un papel importante. Un ejemplo lo encontramos en el caso de "Diana contra el sistema educativo del Estado de California" (1970), en el que 9 niños hispanos de familias granjeras fueron situados en clases de educación especial en base a las puntuaciones obtenidas en un test de inteligencia. Cuando se pasó la misma prueba a los niños, pero esta vez en español, su cociente intelectual aumentó un promedio de 15 puntos, lo que situaba a 7 de ellos por encima del corte establecido para identificar a los niños con necesidad de educación especial. En este caso, el instrumento de medida utilizado estaba sesgado contra el grupo

particular de niños hispanos por su condición de hablantes de inglés no nativos.

El interés por el estudio del sesgo de los tests todavía se vio más intensificado con la aparición del artículo de Jensen (1969) donde defendía el carácter hereditario de la inteligencia. Jensen argumentaba que las diferencias halladas en el cociente intelectual entre blancos y negros eran demasiado grandes para que pudieran ser atribuidas únicamente a diferencias ambientales, y que por lo tanto debían tener un origen genético. La situación discriminatoria descrita anteriormente, unida a la polémica suscitada por el artículo de Jensen, hizo que en ese momento los estudios sobre el sesgo se centraran casi exclusivamente en los instrumentos de medida del cociente intelectual. Este interés se reflejó en la literatura psicométrica de los años 70 con la repentina aparición de gran cantidad de artículos y métodos estadísticos para analizar el sesgo de los tests.

En este contexto, el sesgo de los test es definido como "una fuente de invalidación o de error sistemático en el modo en que un test mide a los miembros de un grupo particular" (Camilli y Shepard, 1994, p. 7). El concepto de grupo es fundamental para poder definir el sesgo de los tests. Los estudios sobre el sesgo de los tests se han hecho en grupos diferenciados por sus características de pertenencia a diferente nivel socioeconómico, raza o etnia, género o cultura. En el contexto de los estudios de comparación entre culturas, el sesgo se operacionaliza como una distorsión sistemática en los resultados para los miembros de uno de los grupos culturales a comparar. Esta distorsión hace que la comparación entre ellos no sea adecuada, e invalida los resultados obtenidos.

En la literatura psicométrica, hablar del sesgo de los tests es prácticamente hablar del sesgo de los items. La mayor parte de los manuales sobre teoría de los tests dedican alguno de sus capítulos a exponer los métodos para el estudio del sesgo de los items (Santisteban, 1990; Martínez Arias, 1994; Muñiz, 1992), e incluso encontramos monografías dedicadas exclusivamente a este tema (Camilli y Shepard, 1994; Holland y Wainer,1993). Sin embargo, desde el marco de los estudios de comparación entre culturas hemos hablado de la amenaza de tres tipos de sesgo: sesgo del constructo, sesgo del método y sesgo de los items. En este capítulo vamos a presentar los diferentes métodos que aparecen en la literatura transcultural para evaluar la presencia de estos sesgos, aunque prestaremos especial atención a los métodos para el estudio del sesgo de los items.

1. PROCEDIMIENTOS PARA EVALUAR EL SESGO DEL CONSTRUCTO

En capítulos anteriores señalábamos la necesidad de evaluar la equivalencia de los constructos en los grupos culturales de interés. Es posible que el mismo constructo sea interpretado y comprendido de forma diferente por dos grupos con culturas diferentes. Por lo tanto, si se va a utilizar un instrumento de medida que ha sido elaborado en uno de los grupos culturales, es importante asegurarse de que el constructo medido por ese instrumento en ese grupo cultural presenta las mismas características en el otro grupo de interés. Es decir, hay que asegurarse de que el instrumento mide adecuadamente el constructo en el nuevo contexto cultural en el que va a ser aplicado.

En el capítulo 2 hemos hablado de los métodos preventivos para evitar el sesgo del constructo. En función de la discrepancia encontrada entre los grupos culturales se debería optar por adaptar el instrumento de medida, o construir uno nuevo. En el caso de que se opte por la adaptación, la administración de un instrumento que no es totalmente idéntico va a complicar los análisis estadísticos de comparación entre culturas. Como veremos más adelante, la teoría de respuesta al ítem o el análisis factorial confirmatorio son técnicas adecuadas ante estos problemas de falta de similitud entre los dos instrumentos. Por otro lado, si la opción es la de construir un nuevo instrumento, los modelos de ecuaciones estructurales (path models) o el análisis de regresión pueden ser técnicas adecuadas para comparar la red nomológica del constructo entre los diferentes grupos culturales.

Pero antes de pasar a los análisis de comparación entre culturas, ¿qué métodos se pueden aplicar para evaluar la no existencia de sesgo del constructo en los datos a comparar? En este caso, los métodos estadísticos sirven de poco. Si el constructo no es idéntico entre los grupos a comparar, los tests estadísticos pueden detectar diferencias entre ellos, pero no ofrecen información para poder comprender la naturaleza de las diferencias encontradas. Por lo tanto, la evaluación del sesgo del constructo debe basarse en un conocimiento profundo de los diferentes grupos culturales y en una búsqueda racional y cualitativa de diferencias en la conceptualización del constructo objeto de estudio.

2. PROCEDIMIENTOS PARA EVALUAR EL SESGO DEL MÉTODO

Van de Vijver y Hambleton (1996) y Van de Vijver y Leung (1996) presentan diferentes procedimientos para examinar la existencia de sesgo del método.

Un procedimiento para analizar el sesgo del método es el uso de matrices monorrasgo-multimétodo. Desde esta aproximación, el constructo psicológico analizado es medido utilizando métodos diferentes. Si las diferencias transculturales encontradas son similares independientemente del método utilizado, ello indicará que no existe sesgo del método. Sin embargo, si las diferencias entre grupos no son las mismas al utilizar diferentes métodos, esto indicará la presencia de sesgo del método.

Otro procedimiento para evaluar el sesgo del método es el análisis test-retest. Este procedimiento es muy útil, sobre todo, en los tests de aptitudes. Si se administra el mismo instrumento en diferentes ocasiones, y se observa un incremento significativo en uno de los grupos en la segunda ocasión, o un patrón de ganancia diferente entre los dos grupos, se puede sospechar la presencia de sesgo del método. En general, si se observa que sujetos de grupos diferentes que han obtenido la misma puntuación en el primer pase del instrumento, obtienen puntuaciones diferentes en la segunda ocasión, se puede poner en duda la validez de los datos recogidos en la primera administración del instrumento. ¿Cómo explicar la diferencia de ejecución en las dos ocasiones? La falta de familiaridad con el procedimiento empleado (tipo de estímulos, formato de repuestas, etc.), puede ser la causa de las diferencias halladas. Los sujetos que carecen de familiaridad con el

método obtienen en la primera ocasión una puntuación inferior a su aptitud real; sin embargo, en la segunda ocasión donde ya han adquirido cierta experiencia con el procedimiento empleado, obtienen una puntuación mayor y más ajustada a su verdadera aptitud. Los sujetos del grupo que estaban familiarizados con el método no sufren diferencias entre los dos pases ya que el método utilizado no había afectado su ejecución en el primer pase.

Otro procedimiento para evaluar el sesgo del método es realizar un estudio piloto en el que se solicite a los examinados todos los comentarios posibles sobre la interpretación de las instrucciones, las alternativas de respuesta, el formato de los items, etc. Este estudio puede ofrecer una información valiosa sobre el grado de familiaridad con los procedimientos empleados, y el nivel de comprensión de las instrucciones a seguir para responder al instrumento de medida.

3. PROCEDIMIENTOS PARA EVALUAR EL SESGO DE LOS ITEMS

Mientras que el sesgo del constructo y el del método implican la idoneidad o no de todo el instrumento para medir el constructo de interés, el sesgo de los items hace referencia a problemas en el instrumento a nivel de los items. Como ya hemos comentado anteriormente, el estudio de este tipo de sesgo y de los métodos para su análisis y detección ha recibido un interés prioritario en la literatura psicométrica, que contrasta con la escasa atención prestada a los otros dos tipos de sesgo. Este hecho es todavía más patente en la actualidad. Para el estudio del sesgo de los tests se han seguido dos aproximaciones estadísticas diferentes: la denominada del sesgo externo, basada en el

paradigma de la validez predictiva, y que se caracteriza por utilizar un criterio externo al test; y la denominada del sesgo interno, que utiliza un criterio interno, normalmente las puntuaciones obtenidas en el test total. Según Osterlind (1979), el sesgo interno se refiere a las propiedades psicométricas de los items. Aunque históricamente surgió antes la aproximación del sesgo externo, en la actualidad predomina la aproximación interna, por lo que el estudio del sesgo de los tests se centra en el análisis de las propiedades psicométricas de sus items.

Antes de pasar a hablar de los procedimientos para evaluar el sesgo de los items resulta necesario hacer una pequeña aclaración terminológica y diferenciar entre dos términos que han sido utilizados con frecuencia como si fueran sinónimos, pero que como vamos a ver a continuación presentan una clara diferencia. Nos referimos a los términos de sesgo y funcionamiento diferencial de los items. Hablar de sesgo y de funcionamiento diferencial de los items en los estudios de comparación entre culturas nos va a llevar a introducir un tercer concepto, la equivalencia. La definición de equivalencia y sus tipos nos permitirá desarrollar un esquema lógico de exposición de los procedimientos para evaluar el sesgo de los items.

La definición dada al principio de este capítulo del concepto de sesgo se ha aplicado indistintamente para hacer referencia al test o a los items. Por ello, podríamos definir el sesgo de los items como "un tipo de invalidación que perjudica a un grupo más que a otro" (Shepard, Camilli, y Averill, 1981). Por otro lado, también podemos decir que un ítem está sesgado cuando sujetos igualmente capaces pero que pertenecen a subgrupos diferentes, presentan diferente probabilidad de responder correctamente al ítem. Ambas definiciones son adecuadas, sin embargo

presentan matices diferentes. La primera recoge el significado social y con matices peyorativos que se ha dado a la palabra sesgo; mientras que la segunda recoge un significado de diferencias estadísticas (Angoff, 1993).

Recordemos de nuevo el contexto en el que se desarrolla el estudio del sesgo: los movimientos reivindicativos por la igualdad de derechos que tienen lugar en Estados Unidos en los años 60. La primera definición encajaría perfectamente en este contexto, ya que recoge el matiz de desventaja e injusticia contra los grupos minoritarios. Como ya hemos visto, para identificar estos items "injustos" se desarrollaron gran cantidad de procedimientos estadísticos. La segunda definición hace referencia a que mediante el uso de esos procedimientos estadísticos se pueden detectar items en los que se observa una probabilidad de ejecución diferente en sujetos de igual capacidad en función del subgrupo al que pertenecen. A partir de ahí sería necesario analizar más profundamente el porqué de ese funcionamiento diferencial del ítem en los dos grupos, utilizando métodos de juicio que permitieran determinar si las diferencias encontradas reflejan realmente la presencia de injusticia en la medida, es decir, si uno de los grupos se ve perjudicado o es injustamente medido debido a las características de ese ítem particular.

Acabamos de introducir un nuevo concepto: el de funcionamiento diferencial de los items (FDI en adelante). La aparición de este término se hizo necesaria para aclarar la confusión generada ante el doble significado con el que se utilizaba la palabras sesgo: el significado social y el significado estadístico (Angoff, 1993; Camilli, 1993; Camilli y Shepard, 1994). De este modo, el término FDI recoge el significado

estadístico y hace referencia a la observación de que un ítem presenta un funcionamiento diferencial en diferentes grupos tras controlar las diferencias en habilidad en ambos grupos.

Algunos autores utilizan los términos sesgo y FDI de forma intercambiable como si fueran sinónimos, y extienden esta confusión terminológica al hablar de los métodos para su estudio. Sin embargo, en base a la diferencia que hemos establecido entre los términos sesgo y FDI, se podría decir que los métodos de detección de FDI (métodos de DFDI, en adelante) son procedimientos empíricos o técnicas estadísticas para identificar aquellos items que presentan un funcionamiento diferencial en distintos grupos. Sin embargo, la obtención de un resultado estadísticamente significativo con un método de DFDI no implica necesariamente que el ítem esté sesgado. Los resultados obtenidos con los métodos de DFDI son una evidencia más en el proceso general de estudio del sesgo de los items. Es necesario realizar posteriormente análisis de juicio para determinar cuáles de los items que han mostrado un funcionamiento diferencial en distintos grupos están realmente sesgados. Un problema de estos métodos es que del mismo modo que un resultado estadísticamente significativo no implica necesariamente la presencia de sesgo, también pueden no detectar la existencia de éste cuando la mayoría de los items están sesgados ya que son métodos que se basan en criterios internos.

El proceso de evaluación del sesgo de los items incluye tanto el uso de métodos de DFDI como de métodos de juicio. Como ya hemos visto, estos últimos pueden utilizarse como un procedimiento a posteriori de búsqueda de explicación al funcionamiento diferencial detectado con los métodos estadísticos. Pero también puede ser utilizado como un

procedimiento a priori en el que jueces expertos intentan identificar aquellos items que crearán una dificultad irrelevante a los miembros de un grupo particular. En general se recomienda el uso conjunto de los dos tipos de métodos. Esto se justifica en base a dos razones. La primera se basa en los resultados de los estudios en los que se ha analizado el grado de acuerdo entre los métodos de juicio y los métodos de DFDI. Estos estudios ponen de manifiesto que por norma general la coincidencia entre la opinión de los jueces y los resultados estadísticos no está por encima de la que se podría dar por puro azar (Plake, 1980; Rengel, 1986; Sandoval y Miille, 1980; Engelhard, Hansche y Rutledge, 1990). Los jueces no tienen demasiado éxito a la hora de predecir los items que presentarán funcionamiento diferencial en dos o más grupos (por ejemplo, hombres versus mujeres, blancos versus negros) detectados por medio de métodos de DFDI. La segunda razón ya ha sido señalada anteriormente: un resultado estadísticamente significativo con un método de DFDI no implica necesariamente la presencia de sesgo. Los métodos de DFDI pueden producir tanto error Tipo I (obtener un resultado estadísticamente significativo en un ítem que no está sesgado), como error Tipo II (no detectar FDI en un ítem que realmente está sesgado). Además, existen otras explicaciones a un resultado significativo en ausencia de sesgo. Los estadísticos FDI detectan multidimensionalidad. Esta multidimensionalidad puede explicar que el ítem presente una dificultad diferente en dos grupos que tienen el mismo nivel en la habilidad primaria o constructo medido por el test, pero que presentan una distribución diferente en una habilidad o factor secundario. Lo que habría que determinar es si las diferencias en la ejecución son debidas a factores secundarios relevantes para el constructo medido, en cuyo caso el FDI estaría indicando multidimensionalidad pero no sesgo; o por el contrario, si el factor

secundario causante de las diferencias en la ejecución es irrelevante para el constructo medido, en cuyo caso sí que se podría hablar de sesgo. Un ejemplo lo encontramos en el trabajo de Shepard et al. (1984), en el que se pasó un test de matemáticas a estudiantes blancos y negros y se observó que todos los items que presentaban una dificultad mayor para los sujetos negros eran problemas que incluían contenido verbal. Este conjunto de items no eran unidimensionales, y para determinar la presencia o no de sesgo era necesario justificar si el factor secundario detectado era relevante o no para el constructo medido. Los autores concluyeron que la aptitud matemática no se reduce únicamente a la habilidad de cómputo numérico, sino que incluye también la capacidad de aplicar esta habilidad numérica a problemas de la vida real que presentan contenido verbal. Por lo tanto, se concluyó que los items en los que se había detectado funcionamiento diferencial (problemas con componente verbal) presentaban multidimensionalidad, pero no estaban sesgados contra el grupo de sujetos negros.

Ya ha sido aclarada la diferencia entre sesgo y FDI. Hemos visto que los métodos estadísticos sirven para detectar la presencia de FDI, y que posteriormente es necesario enjuiciar si existe alguna causa específica que permita etiquetar ese funcionamiento diferencial como sesgo. Las fuentes del sesgo son numerosas, y vienen generadas principalmente por el distinto bagaje cultural, social, económico, etc. Las características socioculturales de los sujetos pueden ser muy diferentes en función del subgrupo al que pertenecen. Esta problemática se acentúa cuando los diferentes grupos no comparten el mismo idioma. En los estudios de comparación entre culturas el proceso de adaptación del instrumento tiene como objetivo evitar el efecto del sesgo de los items, es decir, pretende producir items adaptados a las características

socioculturales y lingüísticas del nuevo grupo cultural, que permitan medir con precisión el constructo de interés. Esto garantiza que si aparecen diferencias entre los dos grupos en un ítem, éstas sean debidas a diferencias reales, y no a la desventaja de uno de los grupos debido al vocabulario utilizado, a problemas de comprensión, etc.

La problemática que surge entonces es la de comprobar que la adaptación ha sido adecuada y que las diferentes versiones del instrumento están midiendo lo mismo. En el contexto de los estudios transculturales suele hacerse referencia a ello mediante el término de equivalencia. Para poder comparar datos procedentes de culturas diferentes, es necesario asegurarse de que esos datos son comparables, o lo que es lo mismo, que han sido recogidos con instrumentos de medida equivalentes. De no existir tal equivalencia, no se contaría con medidas adecuadas para la comparación, ya que los instrumentos estarían midiendo cosas diferentes. Por lo tanto, podemos decir que la equivalencia de los items del instrumento de medida en sus dos versiones, la original y la adaptada, implica que las puntuaciones obtenidas en cada uno de los grupos culturales son comparables. Es posible definir la equivalencia de los items de un test en sus versiones para diferentes lenguas y/o culturas dentro del marco del funcionamiento diferencial de los items: dos versiones de un ítem en diferentes idiomas serán equivalentes si miembros de cada grupo con el mismo nivel en el constructo medido por el test tienen la misma probabilidad de elegir la respuesta correcta si se trata de un ítem de rendimiento, o de elegir la misma alternativa si el ítem forma parte de un cuestionario de personalidad (Hambleton, Swaminathan y Rogers, 1991). Sin embargo, si las probabilidades son diferentes, se podría decir que el ítem presenta funcionamiento diferencial en los dos grupos, y por

lo tanto, que está potencialmente sesgado. Podemos concluir que cuando se habla de la necesidad de utilizar medidas equivalentes, se está haciendo referencia a la necesidad de eliminar el FDI; o lo que es lo mismo, que hablar de items equivalentes es hablar de items que no presentan funcionamiento diferencial en los grupos comparados.

Resumiendo, hemos visto que el funcionamiento diferencial de un ítem en dos grupos diferentes puede ser debido a diferentes causas, y que una de ellas es que las versiones del ítem para los diferentes grupos no sean equivalentes. Conseguir medidas equivalentes del instrumento en diferentes idiomas es una de las preocupaciones centrales en la investigación transcultural (Poortinga, 1983). Por ello, en la literatura transcultural el empleo de métodos de DFDI se ha centrado principalmente en el análisis de la equivalencia de las diferentes versiones del instrumento. A nuestro parecer, el uso preferente que se ha hecho del término de equivalencia resulta también más adecuado porque marca una clara diferencia: los métodos de DFDI en general hacen referencia al estudio del funcionamiento diferencial de un ítem determinado respecto a dos subgrupos diferentes; el término equivalencia hace referencia al funcionamiento diferencial de dos versiones diferentes del mismo ítem utilizadas en grupos culturales diferentes que no comparten el mismo idioma. Vamos a centrarnos en el concepto de equivalencia y en los diferentes métodos utilizados en la literatura transcultural para su evaluación.

Hambleton (1993, 1994, 1996) y Hambleton y Kanjee (1995) presentan una panorámica general de los métodos utilizados para evaluar la equivalencia entre las diferentes versiones del instrumento, y por lo tanto para detectar los items potencialmente sesgados. Estos

autores diferencian entre métodos de juicio (judgmental methods) y métodos estadísticos (statistical methods), y recomiendan siempre que sea posible, utilizar ambos tipos de metodologías en el estudio de la equivalencia. Vamos a pasar a exponerlos a continuación.

3.1. Métodos de juicio

Los métodos de juicio o métodos racionales se basan en la aplicación de un determinado diseño de traducción y en la posterior evaluación de la precisión de dicha traducción por parte de un grupo de traductores o jueces expertos en la materia. Estos deben evaluar si la traducción/adaptación del instrumento ha sido adecuada, y si las diferentes versiones de cada ítem son equivalentes según su opinión.

En la literatura transcultural encontramos dos procedimientos o diseños diferentes para llevar a cabo la traducción del instrumento de medida (Hambleton, 1996; Van de Vijver y Leung, 1996, Hambleton y Kanjee, 1995): la denominada traducción hacia delante o directa, y la traducción hacia atrás o inversa (también llamada "back translation" tomando este término directamente del inglés). A continuación exponemos los pasos que se siguen en cada una de ellas.

3.1.1. Traducción hacia delante o directa

1° Un grupo de traductores traduce el instrumento del idioma fuente al idioma objetivo.

2° Otro grupo de traductores juzga la equivalencia entre las dos versiones del instrumento (la versión fuente y la versión objetivo).

3° Pueden hacerse revisiones de la versión objetivo a partir de los problemas identificados por los traductores tras la comparación de las dos versiones. También puede llevarse a cabo un pase piloto, y en este caso las revisiones de la versión objetivo se hacen a partir de los problemas detectados por los propios sujetos a los que se administra el instrumento.

3.1.2. Traducción hacia atrás o inversa (back translation)

1° Un grupo de traductores traduce el instrumento del idioma fuente al idioma objetivo.

2° Otro grupo de traductores toma el instrumento adaptado y lo vuelve a traducir al idioma fuente.

3° Se compara la versión original del instrumento con la traducción inversa en el mismo idioma, y se hacen análisis racionales sobre su equivalencia. Cuanto mayor sea la similitud entre estas dos versiones, mayor será también la seguridad sobre la equivalencia entre las versiones fuente y objetivo del instrumento.

Hay un acuerdo bastante general en recomendar el uso de la traducción inversa en estudios transculturales; sin embargo, no existe el mismo acuerdo sobre el nivel de utilización actual de este procedimiento. Hambleton (1996) señala que es el método más utilizado, mientras que Wills (1982) recomienda su uso aunque señala que en el campo de la traducción profesional casi nunca es utilizado, ya que

normalmente se recurre a la traducción directa por ser un procedimiento más sencillo.

Sin embargo, a pesar de la profusión y de la recomendación de su uso, se debe ser cuidadoso al utilizar este método. Uno de los problemas de la back translation es que la equivalencia entre la versión original y la traducción inversa puede llevarnos a engaño, ya que no garantiza que la versión traducida para el grupo objetivo sea adecuada (Van de Vijver y Hambleton, 1996). Este procedimiento favorece la traducción literal, por lo que se puede pasar por alto la naturalidad y la facilidad de lectura de la versión objetivo. Es decir, se corre el riesgo de elaborar un instrumento que reproduzca de forma adecuada el lenguaje de la versión original, pero que no resulte fácilmente comprensible en la versión traducida.

El uso de métodos de juicio en los estudios de adaptación de instrumentos de medida es necesario para garantizar su equivalencia, pero no suficiente. Estos métodos proporcionan una información valiosa acerca de la equivalencia de las diferentes versiones del instrumento. Jueces y traductores expertos intentan detectar a priori aquellos items que pueden presentar problemas de equivalencia, lo que permite realizar las modificaciones necesarias para garantizar su adecuado funcionamiento. Sin embargo, el criterio definitivo para evaluar la equivalencia de las versiones del instrumento debe basarse en el análisis de las respuestas dadas por los sujetos a los items del cuestionario. Tal como señalan Hambleton y Kanjee (1995), los sujetos evaluados, a menudo están funcionando en un nivel cognitivo diferente al de los traductores, por lo que es muy posible que una traducción considerada aceptable por los traductores, no lo sea realmente en la

práctica. Además, como ya hemos comentado anteriormente, los jueces no tienen demasiado éxito a la hora de predecir los items que presentarán funcionamiento diferencial en dos o más grupos detectados por medio de métodos estadísticos. Se podría concluir por lo tanto que los items deben probarse sobre el terreno antes de ser utilizados, por lo que los métodos de juicio deberían ser complementados con métodos estadísticos adecuados.

3.2. Métodos estadísticos

Al hablar de métodos estadísticos (o métodos empíricos), Hambleton (1993, 1994, 1996) y Hambleton y Kanjee (1995) diferencian entre diseños y procedimientos. Los primeros hacen referencia al diseño concreto de recogida de datos. Por su parte, los procedimientos hacen referencia a las técnicas estadísticas de análisis que se utilizan posteriormente a la recogida de datos para detectar el funcionamiento diferencial de los items.

3.2.1. Diseños estadísticos

En la literatura transcultural encontramos tres diseños estadísticos diferentes de recogida de datos (Hambleton, 1993; Hambleton, 1994; Hambleton y Kanjee, 1995; Hambleton, 1996) en función de la versión del instrumento administrada (versión original, traducción inversa o versión adaptada), y de las características de la muestra de sujetos a los que se administra el instrumento de medida (monolingües o bilingües).

a) Aplicación de las versiones original y adaptada del instrumento a sujetos bilingües.

La principal ventaja de este tipo de diseño es que se controlan las diferencias individuales (por ejemplo, las diferencias aptitudinales) ya que los mismos sujetos reciben las dos versiones del instrumento. Entre los problemas que plantea podemos citar en primer lugar, la necesidad de comprobar que los sujetos son igualmente capaces en ambos idiomas. Además puede resulta difícil encontrar una muestra grande de personas que dominen a la perfección los dos idiomas (Cziko, 1987; Rosansky, 1979). Por otra parte, puede que los resultados no sean generalizables, ya que las personas bilingües tienden a ser por término medio más capaces que sus homólogos monolingües (Hambleton, 1993).

b) Aplicación de la versión original y de la traducción inversa a sujetos monolingües en el idioma fuente.

Igual que en el diseño anterior, la ventaja de este método es que se controlan las diferencias individuales (por ejemplo, las diferencias aptitudinales). Pero tampoco está exento de problemas. La principal desventaja es que no se recogen datos empíricos sobre la versión del instrumento en el idioma objetivo (versión adaptada del instrumento). Además, el efecto del aprendizaje puede influir sobre el rendimiento de los sujetos. Para evitarlo se pueden utilizar técnicas como el contrabalanceo: se alterna de forma aleatoria el orden de administración para cada sujeto de las dos versiones del instrumento.

c) Aplicación de la versión original a sujetos monolingües en el idioma fuente, y de la versión adaptada a sujetos monolingües en el idioma objetivo.

Este tipo de diseños son los más frecuentemente utilizados (Van de Vijver y Leung, 1996), y según indica Hambleton (1993) es el diseño más útil para el establecimiento de la equivalencia de dos versiones diferentes del instrumento de medida. Algunos ejemplos de su aplicación los encontramos en los trabajos de Ellis (1989, 1991), Ellis y Kimmel (1992), Candell y Hulin (1986), Hulin (1987), y Hulin y Mayer (1986). La principal ventaja de este método es que se utilizan muestras de las poblaciones fuente y objetivo, lo que permite que los hallazgos sobre la equivalencia de las dos versiones del instrumento sean generalizables a las poblaciones de interés. El problema es que este tipo de diseños son incapaces de discriminar entre diferencias culturales y problemas de adaptación del instrumento. Es decir, nos permiten detectar los items que funcionan diferencialmente, pero no nos ofrecen información para determinar si tales diferencias se deben a una adaptación deficiente del instrumento, o a diferencias reales entre las culturas.

3.2.2. Procedimientos estadísticos

Una vez recogidos los datos con cualquiera de los tres diseños estadísticos comentados en el apartado anterior, el siguiente paso sería utilizar los procedimientos estadísticos o técnicas estadísticas adecuadas para identificar aquellos items que no son equivalentes en los grupos comparados. Como ya hemos comentado, el diseño más frecuentemente utilizado es aquel en el que se aplica a sujetos monolingües en el idioma

fuente la versión original y a sujetos monolingües en el idioma objetivo la versión adaptada, por lo que el paso siguiente consistiría en analizar la equivalencia de los items en las dos versiones, la versión original y la adaptada.

Son numerosos los trabajos que han abordado el estudio de la equivalencia de diferentes versiones de cuestionarios traducidos a otras lenguas, y también es diversa la metodología que se ha utilizado. Tras una revisión de estos estudios, llama la atención la amplia variedad terminológica utilizada por los diferentes autores para hacer referencia a la equivalencia, variedad que puede resultar un tanto confusa para el lector. Algunos de los términos utilizados son: equivalencia de las medidas (measurement equivalence), medida equivalente (equivalent measurement), equivalencia psicométrica (psychometric equivalence), equivalencia de la traducción (translation equivalence), equivalencia estructural (structural equivalence), invarianza de las medidas (measurement invariance), invarianza del grupo (group invariance), invarianza de la estructura (structure invariance), invarianza factorial (factorial invariance), sesgo de medida (measurement bias). Ante tal variedad terminológica cabría preguntarse: ¿estos autores están hablando de lo mismo, o hacen referencia a diferentes tipos de equivalencia? Un análisis más detallado pone de manifiesto que el análisis de la equivalencia se realiza a diferentes niveles, lo cual explicaría el uso de diferentes términos para hacer referencia al estudio de diferentes tipos de equivalencia. Sin embargo, la confusión es todavía mayor cuando se observa que se utilizan diferentes términos para referirse a un mismo tipo de equivalencia, y en otras ocasiones se utiliza el mismo término para hablar de tipos distintos de equivalencia.

Nuestra intención es exponer en este apartado los diferentes procedimientos estadísticos que aparecen en la literatura transcultural para el estudio de la equivalencia. Sin embargo, ante este panorama de nomenclaturas variadas y confusas vamos a comenzar tratando el problema de la equivalencia y de los diferentes tipos de equivalencia, para así partir de una clasificación que nos sirva de base para exponer los diferentes procedimientos estadísticos utilizados para su estudio. Las preguntas a responder serían: ¿cuáles son los tipos de equivalencia?, y ¿qué métodos se utilizan para el estudio de cada uno de estos tipos de equivalencia?

Ya hemos comentado anteriormente que la equivalencia de los items de un test en sus diferentes versiones implica que las puntuaciones obtenidas por cada grupo son comparables entre sí. El problema de la equivalencia ha sido ampliamente tratado en la literatura transcultural, y se han identificado diferentes tipos de equivalencia (Berry, 1969; Poortinga, 1971, 1989; Van de Vijver y Poortinga, 1982). Sin embargo, a veces la terminología utilizada ha resultado también un tanto confusa. Por ejemplo, se ha utilizado el término "equivalencia métrica" (metric equivalence) para referirse al caso en que los datos procedentes de dos grupos culturales diferentes presentan propiedades psicométricas similares (Berry, 1969). Este término, interpretado desde una perspectiva psicométrica, debería hacer referencia no a una equivalencia de las propiedades psicométricas, sino a que los datos presentan una unidad de medida común.

Van de Vijver y Leung (1996) señalan también este panorama de terminologías vagas e imprecisas, y proponen una clasificación donde distinguen tres tipos de equivalencia: estructural, de la unidad de

medida, y escalar. La equivalencia estructural (structural equivalence) hace referencia a la semejanza de las propiedades psicométricas de las diferentes versiones de un instrumento adaptado a diferentes culturas. En concreto, la semejanza de las propiedades psicométricas suele hacer referencia a la similitud de la estructura factorial, lo cual alude a que las diferentes versiones del instrumento miden el mismo constructo y lo miden de la misma manera. Para ello, los análisis llevados a cabo para poner a prueba este tipo de equivalencia se basan en el estudio de la similitud del patrón de correlaciones entre los items, y del patrón de correlaciones de los items con el rasgo latente medido; o bien en el estudio de la similitud del patrón de correlaciones de las diferentes versiones del instrumento con otras variables externas. Según la terminología de otros autores, este tipo de equivalencia correspondería a la invarianza de la estructura, o la invarianza a través de grupos (Robie y Ryan, 1996; Brown y Marcoulides, 1996).

Los otros dos tipos de equivalencia hacen referencia a la equivalencia de medida propiamente dicha (measurement equivalence). Se habla de equivalencia de la unidad de medida (measurement unit equivalence) cuando las puntuaciones de los sujetos de las dos culturas comparadas presentan la misma unidad de medida, pero las escalas no tienen un origen común. Un ejemplo ajeno a la medición psicológica que ilustra este caso son las escalas de temperatura en grados Kelvin y grados Celsius. Este tipo de equivalencia impone una serie de restricciones al tipo de comparaciones que se pueden realizar. Cuando se trabaja con diferencias entre puntuaciones, ya sean de diferentes sujetos o del mismo sujeto en diferentes ocasiones, es posible realizar tanto comparaciones intragrupo (dentro de una cultura), como comparaciones entregrupo (comparación transcultural). Sin embargo, cuando se

consideran las puntuaciones en sí mismas, sólo es posible realizar comparaciones intragrupo (entre puntuaciones de sujetos que pertenecen a la misma cultura), pero no será posible realizar una comparación transcultural. Por ejemplo, se ha argumentado que algunos tests de inteligencia únicamente pueden ser aplicados a sujetos del mismo grupo cultural para realizar comparaciones entre sus puntuaciones, pero que la aplicación de estos instrumentos para comparar las puntuaciones de sujetos pertenecientes a distintos grupos culturales no sería válida, ya que la escala presenta diferente origen en cada grupo cultural (Van de Vijver y Leung, 1996).

Por último, se habla de equivalencia escalar (scalar equivalence or full score comparability) cuando las puntuaciones presentan una misma unidad de medida y también un origen común. La equivalencia escalar garantiza que las puntuaciones de los diferentes grupos culturales son comparables, ya que implica que las puntuaciones presentan la misma unidad de medida y también un origen común. Permite tanto la comparación intragrupo (entre puntuaciones de sujetos que pertenecen a la misma cultura) como la comparación entregrupo (entre puntuaciones de sujetos que pertenecen a diferentes grupos culturales), por lo que es el tipo de equivalencia necesario para poder realizar comparaciones transculturales de las puntuaciones de dos sujetos. Ejemplos de variables que poseen equivalencia escalar son el peso o la altura. Sin embargo, en las medidas psicológicas resulta a menudo difícil establecer la equivalencia escalar; es más, por norma general resulta más fácil ofrecer evidencia en contra de la equivalencia escalar, que probar dicha equivalencia (Van de Vijver y Leung, 1996).

Cada uno de los tipos de equivalencia comentados suponen diferentes niveles de similitud entre las medidas por lo que los métodos utilizados para su evaluación van a ser también diferentes. En los estudios de comparación entre culturas el análisis de la equivalencia se ha centrado en la equivalencia estructural y escalar de diferentes versiones de instrumentos traducidos a otros idiomas. Por lo tanto vamos a pasar a exponer los métodos que se han utilizado en la literatura transcultural para el estudio de la equivalencia estructural y escalar, su lógica de funcionamiento y la información que nos proporcionan.

3.2.2.1. Métodos para analizar la equivalencia estructural

Algunas de las técnicas más comúnmente utilizadas para estudiar la equivalencia estructural son el escalamiento multidimensional, y el análisis factorial. En la literatura transcultural encontramos numerosos ejemplos de aplicación de estos métodos, principalmente del análisis factorial exploratorio (Drasgow y Kanfer, 1985; Ben-Porath, Almagor, Hoffman-Chemi y Tellegen, 1995) y del análisis factorial confirmatorio (Robie y Ryan, 1996; Brown y Marcoulides, 1996). Como hemos comentado anteriormente, la terminología utilizada en estos estudios es amplia y variada, y ha originado cierta confusión. Se habla de equivalencia de las propiedades psicométricas, invarianza de las medidas, invarianza de la estructura, e invarianza a través de los grupos. Todos estos términos se utilizan para hacer referencia a un mismo concepto: la equivalencia estructural de las medidas definida por Van de Vijver y Leung.

Hambleton (1993) señala que el estudio de la estructura factorial de las diferentes versiones de un instrumento de medida traducidas a otros idiomas es una herramienta muy útil en el proceso de evaluación de la adecuación de la traducción. Estructuras factoriales similares en los dos grupos proporcionan evidencia de la equivalencia (estructural) de las versiones del test. Por otro lado, estructuras no equivalentes pueden sugerir problemas en el proceso de traducción/adaptación del instrumento. Otro aspecto importante es que si se encuentra la misma estructura factorial en datos procedentes de diferentes grupos culturales, se puede concluir que los constructos psicológicos subyacentes a cada una de las versiones del instrumento son idénticos (Van de Vijver y Leung, 1996), lo cual garantiza la validez de la comparación entre los grupos.

En este apartado vamos a centrarnos en el análisis factorial, más concretamente en el análisis factorial confirmatorio (AFC en adelante) como un método útil, sencillo y adecuado para estudiar la equivalencia estructural de diferentes versiones de un instrumento. Aunque el análisis factorial exploratorio (AFE en adelante) ha sido también muy frecuentemente utilizado para examinar la equivalencia estructural de datos procedentes de grupos culturales diferentes, éste método tienen serias limitaciones (Marsh, 1987a). En primer lugar, no permite definir una estructura factorial determinada, simplemente permite determinar el número total de factores a extraer y el tipo de rotación a aplicar. Además, si la estructura factorial observada no se corresponde con la hipotetizada, no hay ningún procedimiento (por ejemplo, un índice de ajuste) que permita cuantificar en qué medida la estructura hipotetizada se ajusta a los datos. Estas limitaciones se acentúan todavía más cuando pretendemos comparar las estructuras factoriales de dos grupos

diferentes. El AFE no permite definir la estructura factorial en cada grupo, y mucho menos especificar que la estructura debe ser la misma en los dos grupos. Por lo tanto no permite cuantificar en qué medida los datos se ajustan a la estructura hipotetizada en cada grupo, ni tampoco el grado en que las soluciones de cada grupo son similares entre sí. Por todo ello concluimos, junto con otros autores (Alwin y Jackson, 1981; Marsh, 1987a), que el AFE, a pesar de su frecuente uso con este fin, no resulta adecuado como método para analizar la equivalencia estructural.

El uso del AFC sí que permite definir la estructura factorial específica que se quiere poner a prueba, y además ofrece índices cuantitativos del nivel del ajuste de los datos a dicha estructura. Por otro lado, cuando se trabaja con datos procedentes de diferentes grupos culturales, permite realizar una comparación rigurosa de la equivalencia de la estructura factorial de los datos en ambos grupos.

El modelo general que asume el AFC se puede expresar según la siguiente ecuación:

$$X = \Lambda_x \xi + \delta$$

donde,

- X es un vector de rango (q x 1) compuesto por q variables observables,
- Λ_x es una matriz de rango (q x r) compuesta por las saturaciones factoriales que expresan la relación de cada variable observable con la correspondiente variable latente,
- ξ es un vector de rango (r x 1) compuesto por r variables latentes, y
- δ es un vector de rango (q x 1) que representa el error de medición de cada variable observable.

Se ha definido (Λ_X) como la matriz de rango (q x r) de saturaciones factoriales. Otras dos matrices que deben ser definidas son Φ y Θ. La matriz Φ es una matriz de rango (r x r) que contiene las varianzas y covarianzas de las variables latentes; Θ es una matriz de rango (q x q) de covarianzas entre los errores. Poner a prueba una estructura factorial específica supone definir estas matrices, es decir, determinar los parámetros libre y fijos en cada una de ellas, y las relaciones entre ellos. A partir de la ecuación anterior se puede derivar otra ecuación matricial que permite calcular la matriz reproducida de covarianzas entre las variables observables en términos de los parámetros de las matrices Λ_X, Φ , y Θ:

$$\hat{\Sigma} = \Lambda_X \Phi \Lambda_X' + \Theta_\delta$$

Para poner a prueba el modelo factorial hipotetizado se compara la matriz de covarianzas reproducida a partir de los parámetros estimados, con la matriz de covarianzas observada en la muestra. Si la diferencia entre ellas no es estadísticamente significativa, ello indica que el modelo teórico planteado se ajusta de forma adecuada a los datos.

Podemos concluir que el AFC es un método adecuado y útil para analizar la equivalencia estructural de instrumentos de medida utilizados en estudios transculturales. La importancia de confirmar la equivalencia estructural de diferentes versiones de un instrumento utilizadas en contextos culturales diferentes es evidente. Sin embargo, encontrar evidencia de la existencia de equivalencia estructural, no asegura que el origen y la unidad de medida del instrumento sean idénticos en las dos muestras (Van de Vijver y Leung, 1996). Esto es debido a que la equivalencia estructural se basa en la semejanza de correlaciones entre culturas, pero las correlaciones no se ven afectadas

por transformaciones lineales de las variables. Si se multiplican las puntuaciones de los sujetos de una de las culturas por una constante, las correlaciones no se van a ver alteradas, y se obtendrán las mismas saturaciones factoriales. Por tanto, escalas con diferente origen y diferente unidad de medida, pueden presentar estructuras factoriales similares. El estudio de la equivalencia estructural representaría un primer paso en el estudio de la equivalencia de las diferentes versiones del instrumento, previo al estudio de la equivalencia escalar. Vamos pues a pasar a hablar en el siguiente apartado de los métodos para analizar la equivalencia escalar.

3.2.2.2. Métodos para analizar la equivalencia escalar

Según Van de Vijver y Leung (1996), en la literatura se encuentran al menos tres aproximaciones diferentes a la hora de establecer la equivalencia escalar entre puntuaciones de diferentes grupos culturales. En una primera aproximación, se asume la equivalencia de las puntuaciones. Es decir, se administra un test en dos grupos culturales diferentes y se comparan las puntuaciones obtenidas sin realizar previamente ningún análisis para demostrar su equivalencia (Anastasi, 1976; Cattell, 1940; Cattell y Cattell, 1963). Esta aproximación se desaconseja totalmente, y se anima a los investigadores a buscar evidencia empírica de la equivalencia escalar de las puntuaciones.

Las otras dos aproximaciones son procedimientos de validación interna, ya que los datos que se utilizan para analizar la equivalencia son datos recogidos con el mismo instrumento que se está evaluando. La segunda aproximación implica técnicas intraculturales en las que se compara los datos empíricos con hipótesis teóricas para cada cultura. Se

formulan hipótesis sobre el nivel de dificultad o sobre la tasa de aciertos de los items de un instrumento. Por ejemplo, se podrían ordenar los items de un test de aptitudes aritméticas según su nivel de dificultad en base a la complejidad de la operación aritmética requerida. Sería de esperar que las operaciones que requieren la manipulación de números de un sólo dígito sean más fáciles que aquéllas que requieren el uso de números de dos dígitos; también es de esperar que las sumas y restas sean más fáciles que las multiplicaciones y divisiones. Si estas hipótesis no se confirman, ello ofrece evidencia en contra de la equivalencia escalar. Sin embargo, estas técnicas de validación intracultural ofrecen evidencia insuficiente de la presencia de equivalencia escalar (Van de Vijver y Leung, 1996).

La tercera aproximación en el estudio de la equivalencia escalar entre puntuaciones de diferentes grupos culturales, podría ser denominada validación transcultural. El ejemplo más representativo es el estudio del funcionamiento diferencial de los items (Berk, 1982; Holland y Wainer, 1993). En los estudios transculturales el término equivalencia se ha utilizado como sinónimo de ausencia de FDI, señalando que los métodos estadísticos para evaluar la equivalencia de los items serían los métodos de DFDI. Por ejemplo, Hambleton (1993) y Hambleton y Kanjee (1995) citan los procedimientos de la Teoría de Respuesta al Item, el estadístico Mantel-Haenszel, la regresión logística, y el análisis factorial como procedimientos estadísticos para analizar la equivalencia. Esta definición de items equivalentes como items que no presentan funcionamiento diferencial en grupos diferentes recoge un nivel de la equivalencia, la equivalencia escalar. Sin embargo como hemos visto en base a la clasificación de Van de Vijver y Leung el estudio de ésta abarca también otros niveles.

Existen diferentes procedimientos para evaluar la presencia de funcionamiento diferencial de los items. No pretendemos hacer una exposición exhaustiva de todos los métodos de DFDI que aparecen en la literatura psicométrica, sino centrarnos en aquéllos que han sido utilizados mayoritariamente. Hablaremos en primer lugar de los procedimientos basados en la teoría de la respuesta al ítem, pasaremos a desarrollar el procedimiento de Mantel-Haenszel y la regresión logística como procedimientos basados en tablas de contingencia, y finalmente retomaremos el AFC como método que también puede ser utilizado para el análisis del FDI.

Procedimientos basados en la Teoría de la Respuesta al Item

Antes de pasar a hablar de los procedimientos basados en la Teoría de la Respuesta al Item (TRI), es necesario definir una serie de conceptos importantes. En primer lugar, los grupos en los que se estudia el funcionamiento diferencial de los items reciben los nombres de grupo de referencia (el grupo mayoritario), y grupo focal (el grupo minoritario). Otro concepto fundamental es el de curva característica del ítem (CCI). La CCI es la función que relaciona la probabilidad de responder correctamente al ítem con el rasgo o aptitud medida por el test del que forma parte dicho ítem, y se define en base a un conjunto de parámetros. El primer parámetro (a) se refiere a la capacidad de discriminación del ítem; el segundo parámetro (b) se refiere al nivel de dificultad del ítem; y el tercer parámetro (c) se refiere a la probabilidad de acertar el ítem por azar. En la aplicación práctica, los modelos más utilizados son los de dos parámetros, que excluyen el parámetro c. En la figura 3.1. presentamos un ejemplo de CCI. En el eje horizontal, representada con la letra griega theta (θ), aparece la aptitud medida por

el test; en el eje vertical, representada como $P_i(\theta)$, aparece la probabilidad de responder correctamente ese ítem. Puede observarse en la figura que la probabilidad de contestar correctamente al ítem aumenta al aumentar la aptitud, por lo que aquellos sujetos que poseen mayor nivel en la aptitud medida por el test, presentan también mayor probabilidad de dar una respuesta correcta al ítem.

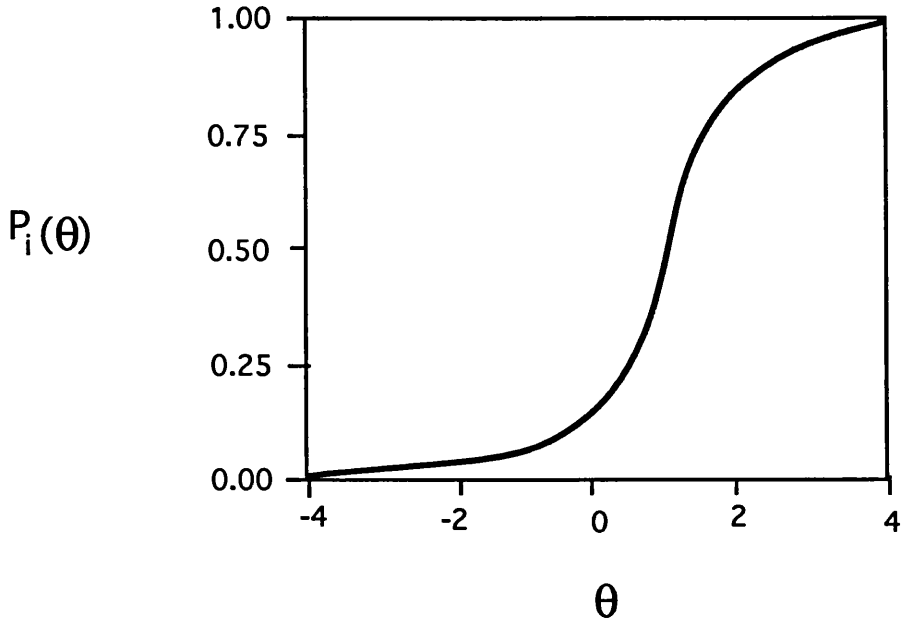


Fig. 3.1. Ejemplo de CCI.

Para evaluar el funcionamiento diferencial de los items se comparan las CCI de los grupos focal y de referencia. Un ítem presenta funcionamiento diferencial si sus parámetros (y por lo tanto su curva característica) difieren de forma significativa en los dos grupos comparados. Ello indica que sujetos con el mismo nivel en la aptitud pero que pertenecen a grupos diferentes, no tienen la misma probabilidad de responder correctamente al ítem. Esto puede verse gráficamente en la figura 3.2. Observamos que si se dibuja una línea vertical desde cualquier nivel de aptitud (θ) del eje horizontal y se prolonga hasta cortar las CCI de los dos grupos, la probabilidad de responder correctamente a ese ítem particular es diferente para cada

grupo. En concreto, el ítem resulta más fácil para el grupo de referencia, ya que para un mismo nivel de aptitud presenta una probabilidad mayor de responder correctamente al ítem.

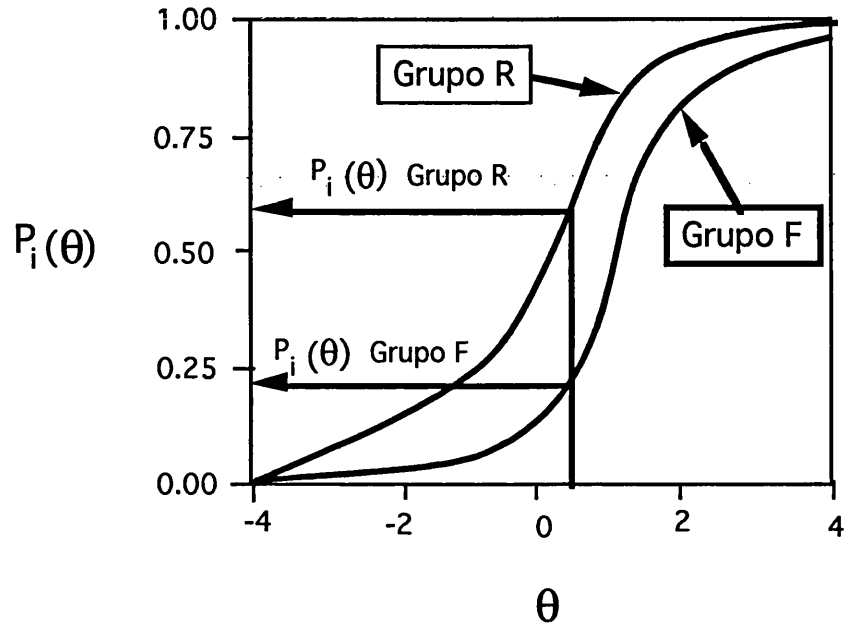


Fig. 3.2. Probabilidad de responder correctamente al ítem de los grupos focal y de referencia para un mismo nivel de aptitud. Ejemplo de FDI uniforme.

Mellenberg (1982) identificó dos tipos de FDI: el uniforme y el no uniforme. El FDI uniforme existe cuando no hay interacción ente el nivel de aptitud o habilidad y el grupo de pertenencia, es decir, cuando la probabilidad de responder al ítem correctamente es mayor para uno de los grupos en todos los niveles de aptitud. La figura 3.2. representa un ejemplo de FDI uniforme o consistente. En términos de la TRI, este tipo de FDI ocurre cuando las CCI de los grupos focal y de referencia son diferentes pero no se cruzan, es decir, son paralelas. Por lo tanto, uno de los grupos presenta mayor probabilidad de responder correctamente al ítem en todos los niveles de aptitud. Este es el caso que se da cuando las dos CCI tienen el mismo parámetro a, pero el parámetro b es diferente.

El FDI no uniforme existe cuando hay interacción entre el nivel de aptitud y el grupo de pertenencia, es decir, cuando la diferencia en la probabilidad de dar una respuesta correcta al ítem para los dos grupos no es la misma para todos los niveles de aptitud. En términos de la TRI, el FDI no uniforme o inconsistente ocurre cuando las CCI de los grupos focal y de referencia son diferentes y además se cruzan en algún punto de la escala de aptitud. En este caso, los parámetros a y b son diferentes en las dos CCI. La figura 3.3. representa un ejemplo de este tipo de FDI. Como puede observarse, para niveles bajos de aptitud los sujetos del grupo focal presentan mayor probabilidad de responder correctamente al ítem; sin embargo, para niveles altos de aptitud son los sujetos del grupo de referencia los que presentan mayor probabilidad de éxito en el ítem.

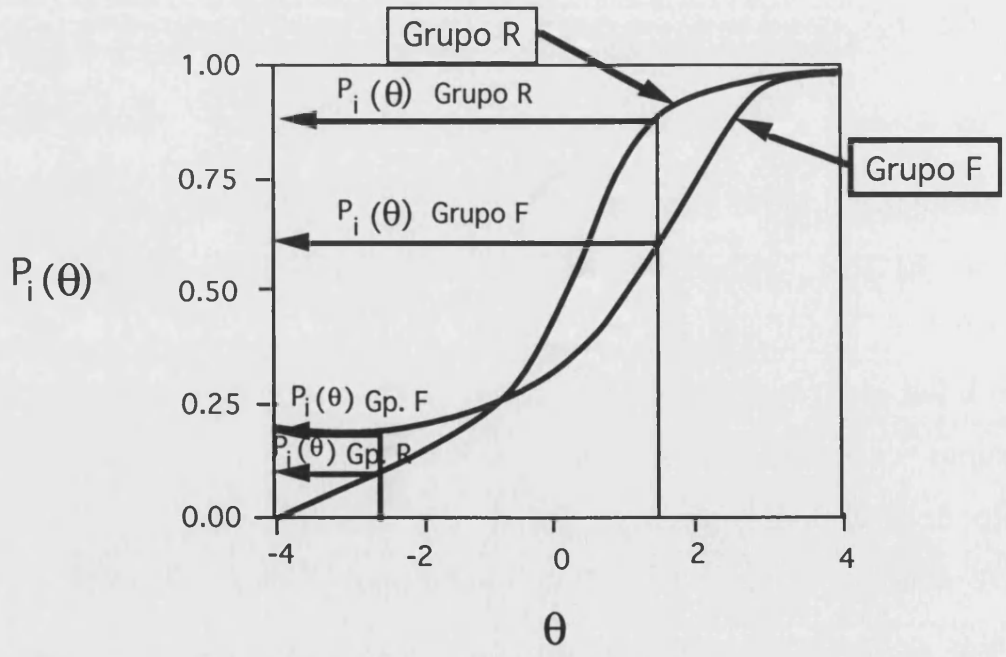


Fig. 3.3. Probabilidad de responder correctamente al ítem de los grupos focal y de referencia. Ejemplo de FDI no uniforme.

Hemos visto que un ítem presenta FDI si los parámetros de las CCI en los grupos focal y de referencia son diferentes. Para analizar el funcionamiento diferencial del ítem se deben estimar los valores de los

parámetros en cada uno de los grupos; pero antes de realizar las comparaciones es necesario transformar los parámetros a una métrica común. Existen dos métodos para obtener parámetros de los items comparables en los dos grupos (Camilli y Shepard, 1994): el método de muestras separadas (separate sample method), y el método del test de anclaje (anchor test method). En el método de muestras separadas los parámetros del ítem son estimados de forma independiente para cada grupo. La posterior transformación de estos parámetros a una métrica común puede realizarse mediante diferentes métodos que pueden ser agrupados en tres grandes categorías: métodos basados en los momentos, métodos basados en la curva característica, y otros métodos no clasificables en ninguna de las categorías anteriores (Navas Ara, 1996). Uno de los métodos más utilizados y que se incluiría dentro de la segunda categoría, es el método de la curva característica del test, más conocido como método de Stocking-Lord, ya que aunque fue formulado inicialmente por Haebara (1980), posteriormente fue reformulado por Stocking y Lord (1993). En el método del test de anclaje, la estimación de los parámetros se hace de forma simultánea en los dos grupos, por lo que la transformación a una métrica común se realiza también al mismo tiempo que la estimación de los parámetros. En este método, los items del test son divididos en dos conjuntos: el conjunto de items a analizar, y el conjunto de items de anclaje. Lo más habitual es que el conjunto de items a analizar esté formado por un único ítem y que el conjunto de items de anclaje esté formado por el resto de items que componen el test. Por lo tanto, el análisis de FDI se hace ítem a ítem, utilizando el resto de items que componen el test como items de anclaje que permiten establecer una métrica común. Durante el proceso de estimación los dos conjuntos de items son tratados de forma diferente: los parámetros de los items de anclaje son constreñidos a ser idénticos en los dos grupos,

mientras que los parámetros del ítem a analizar pueden tomar valores diferentes en los dos grupos.

Quedan todavía dos preguntas por responder: ¿cómo se mide el FDI cuando se utilizan los procedimientos basados en la TRI?, y ¿qué procedimientos se utilizan para determinar la significación estadística del FDI detectado? Camilli y Shepard (1994) distinguen ente métodos para medir el FDI y métodos para evaluar la significación estadística del FDI. Los primeros miden la magnitud o el tamaño del FDI, pero tienen el problema de que no suelen disponer de contrastes para determinar su significación estadística. Los índices del área son un ejemplo de métodos para medir el FDI. Estos índices se basan en la medida del área entre las dos curvas (ver figura 3.4.), es decir, cuantifican el FDI en función del área existente entre las CCI de los grupos comparados, de tal forma que cuanto mayor es el área entre las dos curvas, mayor es la magnitud del FDI.

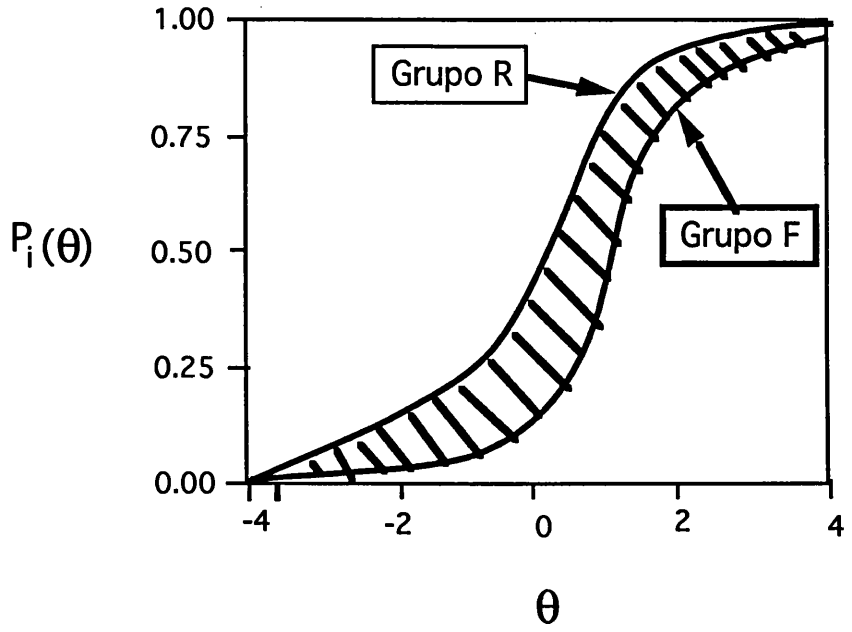


Fig. 3.4. Representación visual de la magnitud del FDI. La medida del área entre las dos curvas (área rayada) permite cuantificar la magnitud del FDI.

Los métodos para evaluar la significación estadística del FDI se basan en la comparación de los parámetros de los items, y entre otros podemos citar el contraste de las diferencias en b, el estadístico chi-cuadrado de Lord (Lord, 1980), la distribución muestral empírica de los índices FDI, y las medidas basadas en la comparación de modelos. Queda fuera del propósito de este trabajo exponer cada uno de estos métodos, por lo que remitimos al lector interesado al trabajo de Camilli y Shepard (1994).

Aplicando todo lo visto hasta el momento al contexto de los estudios transculturales, podemos afirmar que un ítem es equivalente a su versión traducida si sus curvas características en ambos grupos son iguales, es decir, si tienen los mismos parámetros. Sujetos que pertenecen a diferentes culturas o que hablan diferentes idiomas, pero que tienen el mismo nivel en el rasgo latente, presentarán la misma probabilidad de responder correctamente al ítem (o a su traducción equivalente). Las escalas compuestas por items no equivalentes pueden ofrecer medidas sesgadas, por lo tanto dichos items deberían ser sometidos a análisis de juicio con el objeto de determinar si son adecuados para realizar comparaciones transculturales.

Los procedimientos de la teoría de respuesta al ítem han sido ampliamente utilizados en el estudio transcultural de diferentes constructos: autoconcepto (Leung y Drasgow, 1985), satisfacción laboral (Hulin y Mayer, 1986), inteligencia (Ellis, 1989; Van de Vijver, 1988), personalidad (Ellis, Becker y Kimmel, 1993), y actitudes hacia la salud mental (Ellis y Kimmel, 1992). El procedimiento estándar para la aplicación de la TRI en el estudio del funcionamiento diferencial de los items en estudios transculturales es el siguiente:

1°. Uno de los supuestos que debe cumplirse para poder aplicar los métodos de la TRI es que los items evaluados representen un único rasgo latente, o lo que es lo mismo, que el conjunto de items sea unidimensional. La comprobación de la unidimensionalidad se lleva a cabo realizando análisis factoriales de forma separada con los datos de cada grupo cultural. En el caso de que la escala sea multidimensional, la unidimensionalidad de cada subescala se evalúa de forma independiente. Por otro lado, diferentes autores (Drasgow y Parsons, 1983; Harrison, 1986) presentan evidencia de que ciertos modelos de la TRI, incluyendo los modelos de dos y tres parámetros, son capaces de tolerar alguna violación del supuesto de la unidimensionalidad.

2°. Para la estimación de los parámetros suele utilizarse el método de muestras separadas, por lo que tras elegir un modelo de TRI adecuado a los datos a analizar, se estiman los parámetros de los items en las dos muestras en base al modelo elegido.

3°. Los parámetros identificados para cada grupo cultural son transformados a una métrica común por medio de procedimientos de ajuste iterativo (el procedimiento más comunmente utilizado es el método de Stocking-Lord).

4°. Se procede a la comparación de los parámetros obtenidos y equiparados en las dos muestras. Los items que presentan funcionamiento diferencial son detectados y eliminados en base a los resultados de contrastes estadísticos para evaluar la significación del FDI. Uno de los métodos que se utiliza con más frecuencia es el estadístico chi-cuadrado de Lord.

5° Se estima de nuevo los parámetros, pero esta vez incluyendo solamente aquellos items que no habían presentado funcionamiento diferencial. Los parámetros son transformados a una métrica común por el mismo procedimiento de ajuste, y se repite el proceso de comparación. Si se vuelven a detectar items que presentan FDI, éstos son eliminados, y se vuelve a repetir todo el proceso hasta que ya no se detectan items con FDI.

6°. El funcionamiento diferencial de un ítem puede ser debido a problemas en la traducción o a diferencias culturales. Esto debe ser evaluado mediante un análisis de contenido de los items que han presentado FDI. Los items que han presentado funcionamiento diferencial debido a una traducción inadecuada pueden ser corregidos y volver a ser evaluados. Si la corrección del ítem elimina su funcionamiento diferencial, el ítem puede ser integrado de nuevo en la escala, lo que permite mantener la integridad del test original. Si la explicación del FDI es por diferencias culturales, el ítem no podrá ser utilizado en la comparación transcultural, pero sí que nos ofrecerá información valiosa sobre las diferencias culturales entre los grupos analizados.

La TRI tiene una serie de características que la hacen apropiada para su aplicación en estudios transculturales. La estimación de los parámetros de los items no depende del nivel del grupo analizado. Esta característica no se da en los procedimientos de la teoría clásica de tests, ya que en ésta la dificultad de un ítem sí depende del nivel de habilidad de los sujetos del grupo. De forma similar, en la TRI la estimación del nivel de habilidad es independiente de los items del instrumento. Sin embargo, la TRI también presenta una serie de limitaciones. En primer

lugar, la aplicabilidad de los modelos de la TRI puede verse reducida debido a los supuestos tan estrictos que deben ser considerados, particularmente en el modelo de un parámetro o modelo de Rasch (Van de Vijver y Leung, 1996). Por otro lado, se requieren grandes tamaños muestrales para asegurar la obtención de estimaciones estables, especialmente en el modelo de tres parámetros. Aunque el modelo de Rasch requiere muestras de 200 sujetos para cada grupo, en general, los otros modelos de la TRI necesitan muestras más numerosas (Hambleton, 1996).

Esta relación entre la potencia del modelo y el tamaño muestral genera importantes limitaciones cuando el tamaño de la muestra no es muy grande (Camilli y Shepard, 1994). En situaciones en las que se dispone de una muestra pequeña, solamente pueden aplicarse modelos restrictivos (como por ejemplo el modelo de Rasch) ya que éstos requieren un menor tamaño muestral para la estimación de los parámetros. Pero precisamente, cuando el tamaño muestral es pequeño, resulta difícil evaluar el ajuste del modelo. Por lo tanto, la precisión de los resultados del análisis de FDI dependerá en gran medida de la validez del modelo de TRI elegido por el investigador. Si el modelo de un parámetro (elegido en base al reducido tamaño muestral del que se dispone) presenta una ajuste adecuado a los datos, entonces se puede confiar en la precisión de los resultados del análisis de FDI. En el caso de que el ajuste del modelo no sea adecuado, los resultados no serán tan precisos. Pero en ambos casos, el ajuste del modelo no puede ser evaluado con precisión, debido al reducido tamaño de la muestra.

En general, cuando se cuenta con muestras lo suficientemente grandes como para poder evaluar adecuadamente el ajuste del modelo,

se recomienda el uso de métodos basados en la TRI, ya que proporcionan resultados más generalizables. Sin embargo, cuando el tamaño muestral es pequeño, o cuando se prefiere un tipo de análisis menos complejo y más rápido, se puede recurrir a otros procedimientos, como los que vamos a ver en el siguiente apartado.

Procedimientos basados en tablas de contingencia

Otra aproximación al estudio del FDI que guarda ciertas semejanzas con los modelos de la TRI, pero que no presenta sus limitaciones, son los procedimientos basados en tablas de contingencia (TC en adelante). Estos procedimientos son más sencillos en su cálculo, y permiten un ahorro temporal considerable en comparación con los procedimientos basados en la TRI que resultan más complejos y requieren más tiempo para su aplicación. Además, los procedimientos basados en las TC requieren tamaños muestrales menores, y son más fáciles de comprender, resultando más asequibles a aquellas personas que no tienen amplios conocimientos psicométricos.

Tanto los procedimientos basados en la TRI como los basados en las TC analizan el FDI para dos grupos comparando la ejecución de sujetos que tienen el mismo nivel de aptitud. Desde la aproximación de las TC se utiliza la puntuación total en el test (operacionalizada como número total de aciertos) para determinar niveles de aptitud comparables. Otra característica común es que las dos aproximaciones se aplican de forma iterativa. Se puede decir que los procedimientos basados en las TC son procedimientos no paramétricos, ya que no se basan en modelos de medida explícitos que proporcionen estimaciones de parámetros como sí sucede en los procedimientos basados en la TRI.

El nombre que reciben estos procedimientos hace referencia al modo en que los datos pueden ser tabulados. El primer paso en el análisis del FDI consiste en codificar los datos en tablas de contingencia de tres vías. Para ello se necesita conocer la siguiente información de cada sujeto:

1° Grupo de pertenencia: si pertenece al grupo focal o al de referencia.

2° Respuesta dada a cada uno de los items: codificada como acierto (=1) o error (=0).

3° La puntuación total en el test: que se obtiene sumando el número total de items contestados correctamente. Para un test formado por k items, la puntuación total puede tomar valores entre 0 y k, por lo que se pueden establecer k+1 niveles de puntuación.

Se agrupa a los sujetos por niveles de puntuación en el test total, y para cada ítem se construyen k+1 tablas que recogen la información en ese ítem de los sujetos agrupados por niveles de puntuación total. Es decir, se construyen un total de k (número de items) por k+1 (número de niveles de puntuación) tablas, que recogen la siguiente información:

1° Grupo de pertenencia: grupo focal o de referencia.

2° Respuesta dada a ese ítem particular (acierto o error).

3° Nivel de puntuación total en el test de los sujetos a los que pertenecen los datos de la tabla.

Veamos un ejemplo. Supongamos un test formado por 8 items. Se puede agrupar a los sujetos en 9 niveles en base a su puntuación total en el test. El "nivel 0" reuniría a todos los sujetos que no han acertado

ningún ítem, el "nivel 1" a los que han acertado solamente un ítem, y así sucesivamente hasta el "nivel 8" que agruparía a los sujetos que han contestado correctamente los 8 items del test. Por lo tanto, para el análisis de cada ítem tendríamos un total de 9 tablas de contingencia. La figura 3.5. sería un ejemplo de una tabla 2x2 (grupo por respuesta en el ítem) que recoge la información del ítem i en el nivel j de aptitud (puntuación total en el test). Para completar la clasificación en una tabla de contingencia de tres vías sería necesario construir otras 8 tablas semejantes a las del ejemplo que recogieran la información de los sujetos de los otros niveles de aptitud.

Nivel j (j=0, 1, 2, ..., K)

Grupo	Puntuación en el ítem i: 1 (acierto)	0 (error)	Total
R	8	2	10
F	6	4	10
Total	14	6	20

Fig. 3.5. Ejemplo de tabulación de datos mediante tablas de contingencia.

La tabla recoge la información de 20 sujetos que tienen la misma puntuación total. Puede observarse que de los sujetos del grupo de referencia 8 lo acertaron y 2 lo fallaron, mientras que de los sujetos del grupo focal 6 lo acertaron y 4 lo fallaron.

La notación general de la tabla (Holland y Thayer, 1988) aparece representada en la figura 3.6.

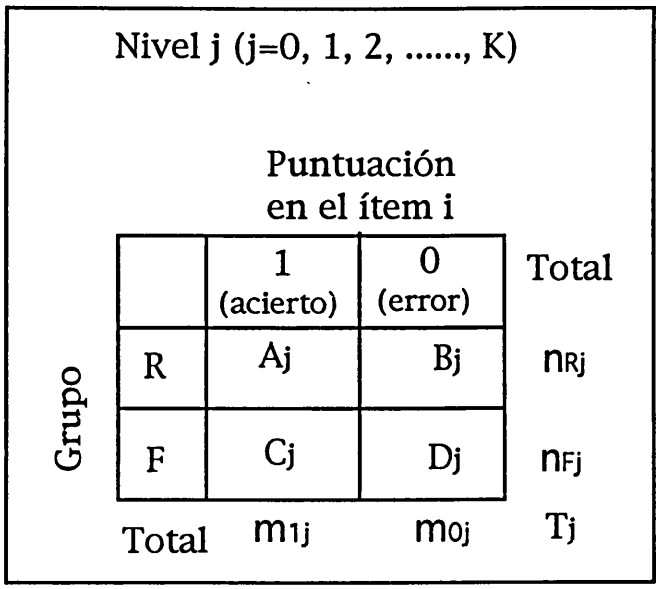


Fig. 3.6. Notación general de la tabla de contingencia 2X2 (Holland y Thayer, 1988).

Ya hemos comentado que para completar la clasificación en una tabla de contingencia de tres vías sería necesario construir k+1 tablas 2X2, una para cada nivel de puntuación total. Sin embargo, en el "nivel 0" los sujetos han fallado todos los items, y en el "nivel 1" los sujetos han acertado todos los items, por lo que en ambos casos encontraremos el mismo patrón de respuesta al ítem en el grupo focal y en el de referencia. En la práctica, las tablas de estos dos niveles se desestiman, construyéndose un total de k-1 tablas para cada ítem.

¿Cuál es la lógica para el estudio del FDI en los procedimientos que se basan en estas tablas de contingencia? Los sujetos del grupo focal y del grupo de referencia de cada tabla tienen la misma puntuación total en el test, lo que se compara es la proporción de respuestas correctas en ese ítem de los sujetos de cada uno de los grupos. Se puede establecer un paralelismo entre los procedimientos basados en la TRI y los basados en las TC. En la TRI se establece la probabilidad de aceptar el ítem $P_i(\theta)$ en función del nivel de aptitud estimado. En las tablas de contingencia se establece la proporción de aciertos en el ítem en función del nivel en la

aptitud determinado mediante la puntuación total obtenida en el test. Por lo tanto, la proporción estimada de respuestas correctas puede interpretarse como la probabilidad estimada de éxito en ese ítem.

La proporción de aciertos en cada grupo se obtiene dividiendo el número de personas que responden correctamente al ítem por el número total de personas del grupo en ese nivel. Por ejemplo, la proporción de aciertos en el grupo de referencia (p_{Rj}) se obtiene dividiendo A_j/n_{Rj}. Del mismo modo, la proporción de aciertos en el grupo focal (p_{Fj}) se obtiene dividiendo C_j/n_{Fj}. Por otro lado, la proporción de errores en cada grupo se obtiene dividiendo el número de personas que fallan el ítem por el número total de personas del grupo en ese nivel. Por ejemplo, la proporción de errores en el grupo de referencia (q_{Rj}) se obtiene dividiendo B_j/n_{Rj}. Del mismo modo, la proporción de errores en el grupo focal (q_{Fj}) se obtiene dividiendo D_j/n_{Fj}. La figura 3.7. recoge la notación general de una tabla de contingencia de las proporciones de aciertos y errores para los sujetos de los grupos focal y de referencia en un determinado nivel de aptitud.

Nivel j (j=0, 1, 2, ..., K)

Grupo	Puntuación en el ítem i: 1 (acierto)	0 (error)	Total
R	p_{Rj}	q_{Rj}	1
F	p_{Fj}	q_{Fj}	1

Fig. 3.7. Notación general de la tabla de contingencia de las proporciones de aciertos y errores para los sujetos del nivel j.

Si tomamos los datos del ejemplo de la figura 3.5. y calculamos los valores de las proporciones de aciertos y errores para cada uno de los grupos, podríamos representarlos como aparece en la figura 3.8.

Nivel j (j=0, 1, 2, ..., K)

Puntuación en el ítem i

Grupo	1 (acierto)	0 (error)	Total
R	0.8	0.2	1
F	0.6	0.4	1

Fig. 3.8. Tabla de las proporciones de aciertos y errores para los datos del ejemplo de la figura 3.4.

Son muchos los procedimientos estadísticos que a partir de los datos de las TC comparan las proporciones de aciertos de los dos grupos para evaluar la existencia de FDI. Sin embargo no vamos a entrar a explicar cada uno de ellos (remitimos al lector interesado a Camilli y Shepard, 1994). Hablaremos de los dos más frecuentemente utilizados: la regresión logística y el procedimiento de Mantel-Haenszel.

El procedimiento de Mantel-Haenszel fue desarrollado por Mantel y Haenszel (1959) y aplicado al estudio del FDI por Holland (1985) y más tarde por Holland y Thayer (1988).

Estos autores definen α_{MH} como una medida de FDI

$$\alpha_{MH} = \frac{\sum_{j=1}^{k} A_j D_j / T_j}{\sum_{j=1}^{k} B_j C_j / T_j}$$

donde Aj, Bj, Cj, Dj y Tj tienen el significado que hemos visto en la figura 3.5. Un valor de 1 significa que el ítem presenta la misma proporción de aciertos en ambos grupos, y por lo tanto indica que no hay FDI. Un valor mayor que 1 indica que los sujetos del grupo de referencia presentan una proporción mayor de respuestas correctas al ítem, por lo tanto indicaría presencia de FDI a favor del grupo de referencia. Finalmente, un valor menor que 1 indica que los sujetos del grupo focal presentan una proporción mayor de respuestas correctas al ítem, por lo tanto indicaría presencia de FDI, en este caso a favor del grupo focal.

Si se calcula el logaritmo neperiano de esta medida de FDI se obtiene un índice con signo:

$$\beta_{MH} = \ln \alpha_{MH}$$

Valores positivos de β_{MH} indican presencia de FDI a favor del grupo de referencia, y valores negativos indican presencia de FDI a favor del grupo focal. Un valor $\beta_{MH=0}$ indica ausencia de FDI.

Tanto α_{MH} como β_{MH} indican el tamaño del FDI, pero no su significación estadística. Para ello, Mantel y Haenszel (1959) propusieron

un estadístico chi-cuadrado que permite contrastar las hipótesis siguientes:

H_0: $\alpha_{MH} = 1$ $\qquad$ H_0: $\beta_{MH} = 0$

o bien

H_1: $\alpha_{MH} > 1$ $\qquad$ H_1: $\beta_{MH} \neq 0$

El citado estadístico de contraste sigue una distribución χ^2 con un grado de libertad, y se define según la siguiente expresión:

$$\chi^2_{MH} = \frac{\left\{ \left| \sum_{j=1}^{k} A_j - \sum_{j=1}^{k} E(A_j) \right| - 0.5 \right\}^2}{\sum_{j=1}^{k} VAR(A_j)}$$

donde:

$$VAR(A_j) = \frac{n_{Rj}\, n_{Fj}\, m_{1j}\, m_{0j}}{T_j^2 (T_j - 1)}$$

$$E(A_j) = \frac{n_{Rj}\, m_{1j}}{T_j}$$

El procedimiento de Mantel-Haenszel es actualmente uno de los más utilizados para evaluar el funcionamiento diferencial de los items de respuesta dicotómica (Van de Vijver y Leung, 1996), aunque también se han propuesto extensiones de este procedimiento para analizar el FDI

de items politómicos (Zwick, Donoghue y Grima, 1993; Welch y Miller, 1995; Welch y Hoover, 1993). Su amplia difusión puede ser debida a que es un método fácil de aplicar y rápido de ejecutar. Por ejemplo, Swaminathan y Rogers (1990) señalan que el procedimiento basado en el modelo de regresión logística requiere de 3 a 4 veces más tiempo de trabajo de cálculo por parte del ordenador que el procedimiento Mantel-Haenszel, y señalan que los procedimientos más costosos en cuanto a tiempo de cálculo se refiere son los procedimientos basados en la TRI. Otra de las ventajas del procedimiento Mantel-Haenszel es que no requiere tamaños muestrales demasiado grandes. Para aplicar esta técnica se necesita que la muestra tenga un mínimo de 200 sujetos en cada población, mientras que en otros procedimientos, como los basados en la TRI, se necesitan tamaños muestrales mucho mayores.

Una de las mayores críticas que ha recibido este procedimiento ha sido su insensibilidad para detectar FDI no uniforme. Diferentes autores han puesto de manifiesto que el estadístico Mantel-Haenszel está diseñado para detectar FDI uniforme, y que por lo tanto es prácticamente insensible a la detección del FDI no uniforme. Encontramos ejemplos de ésta crítica en trabajos como el de Hambleton y Rogers (1989) en el que se detectó FDI no uniforme en un conjunto de items utilizando procedimientos basados en la TRI; sin embargo el estadístico Mantel-Haenszel fue incapaz de detectar este tipo de FDI. Resultados un poco más esperanzadores se encontraron en los trabajos de Swaminathan y Rogers (1990) y Rogers y Swaminathan (1993), en los que se comparó la efectividad para detectar FDI uniforme y no uniforme de los procedimientos basados en la regresión logística y del estadístico Mantel-Haenszel. Los resultados a los que llegaron estos estudios fueron que los dos procedimientos resultaban igualmente potentes para

detectar el FDI uniforme, pero que la regresión logística era considerablemente más potente que el estadístico Mantel-Haenszel en la detección del FDI no uniforme. Estos autores concluyeron que el procedimiento de Mantel-Haenszel puede detectar satisfactoriamente el FDI no uniforme en tests que son muy fáciles o muy difíciles, pero cuando el test presenta una dificultad moderada, el estadístico Mantel-Haenszel falla en la detección del FDI no uniforme incluso cuando éste es muy evidente. A pesar de que estos resultados ya no permiten mantener la insensibilidad absoluta del procedimiento de Mantel-Haenszel para detectar el FDI no uniforme, los citados autores, desde una postura crítica, recomendaban utilizar procedimientos alternativos que fueran capaces de detectar adecuadamente los dos tipos de FDI.

Los resultados de éstos trabajos parecían apuntar hacia una seria deficiencia del procedimiento de Mantel-Haenszel que podía eclipsar todas las ventajas de simplicidad de cálculo y reducido tamaño de la muestra requerida que se le habían atribuido. Sin embargo, trabajos posteriores han puesto de manifiesto que esta crítica tiene sus limitaciones y que una simple modificación del procedimiento estandard de aplicación convierte al estadístico de Mantel-Haenszel en un método eficaz para detectar el FDI no uniforme (Mazor, Clauser, y Hambleton, 1994). Hemos visto anteriormente que el procedimiento de Mantel-Haenszel aplicado de forma estandard permite detectar el FDI no uniforme en items muy fáciles o muy difíciles; sin embargo, no es capaz de detectar items que presentan FDI no uniforme cuando éstos tienen un índice de dificultad medio. Mazor et al. (1994) introdujeron una pequeña modificación sobre el procedimiento estándar: dividieron la muestra total en dos grupos en función de la puntuación obtenida en el test, de tal forma que uno de los grupos estaba formado por los sujetos que

habían obtenido una puntuación en el test por encima de la media, y el otro grupo estaba formado por los sujetos que habían obtenido una puntuación por debajo de la media. Posteriormente calcularon el estadístico de Mantel-Haenszel de forma independiente en cada uno de estos dos grupos, y demostraron que de esta forma se incrementaba considerablemente la tasa de identificación de items con FDI no uniforme con respecto a la tasa obtenida con el procedimiento estándar, sin que esto supusiera un aumento de la tasa de error Tipo I. Posteriormente, distintos trabajos de simulación que han utilizado esta variación del procedimiento de Mantel-Haenszel para detectar FDI no uniforme, han puesto de manifiesto que representa una alternativa eficaz y adecuada para la detección de este tipo de FDI (Fidalgo, 1996; Fidalgo y Mellenbergh, 1995).

El modelo estándar de regresión logística para predecir una variable dependiente dicotómica a partir de otras variables independientes (Bock, 1975) ha sido aplicado por Swaminathan y Rogers (1990) para detectar FDI. En base a éste modelo se puede definir la siguiente ecuación para predecir la probabilidad de responder correctamente a un ítem:

$$P(u = 1/\theta) = \frac{e^{(\beta_0+\beta_1\theta)}}{\left(1+e^{(\beta_0+\beta_1\theta)}\right)}$$

donde u es la respuesta al ítem, θ es la aptitud observada de un sujeto, β_0 es el parámetro intercepto (intecept parameter), y β_1 es el parámetro de la pendiente (slope parameter).

El análisis de FDI supone la comparación del funcionamiento del ítem en dos grupos diferentes, por lo que la ecuación anterior puede ser desarrollada para cada uno de los grupos:

$$P(u_{ij} = 1/\theta_{ij}) = \frac{e^{(\beta_{0j}+\beta_{1j}\theta_{ij})}}{\left(1+ e^{(\beta_{0j}+\beta_{1j}\theta_{ij})}\right)} \quad , i = 1, ..., n_j \quad j = 1, 2$$

En esta ocasión, u_{ij} representa la respuesta al ítem de la persona i en el grupo j, β_{0j} es el parámetro intercepto para el grupo j, β_{1j} es el parámetro de la pendiente para el grupo j, y θ_{ij} es la aptitud del sujeto i del grupo j.

Un ítem muestra FDI si sujetos con el mismo nivel de aptitud en la variable medida pero que pertenecen a grupos diferentes, no tienen la misma probabilidad de responder correctamente al ítem. Expresado en términos del modelo de regresión logística, se podría concluir que un ítem presenta FDI si sus curvas de regresión logística para los dos grupos no son iguales, o lo que es lo mismo, si los parámetros que definen estas curvas no son iguales.

Se ha hablado de FDI uniforme cuando no hay interacción entre el nivel de aptitud y el grupo de pertenencia. Expresado en términos del modelo de regresión logística se puede decir que hay FDI uniforme cuando las curvas de regresión en los dos grupos son paralelas (tienen la misma pendiente), pero no coinciden. Es decir, cuando $\beta_{11}=\beta_{12}$, pero $\beta_{01}\neq\beta_{02}$. Por otro lado, se ha hablado de FDI no uniforme cuando sí que existe interacción entre el nivel de aptitud y el grupo de pertenencia. En

términos de la regresión logística, cuando las curvas de regresión en los dos grupos no son paralelas, es decir, no tienen la misma pendiente, se habla de FDI no uniforme. En este caso, $\beta_{01}=\beta_{02}$, pero $\beta_{11}\neq\beta_{12}$.

Una forma alternativa pero equivalente de expresar el modelo es la siguiente:

$$P(u=1)=\frac{e^{z}}{(1+e^{z})}$$

donde: $z=\tau_0+\tau_1\theta+\tau_2 g+\tau_3(\theta g)$

En esta formulación, θ representa el nivel observado del sujeto en el rasgo o aptitud medido por el test. La variable g representa el grupo de pertenencia, que se define designando de forma arbitraria valores diferentes al grupo focal y al grupo de referencia. Por ejemplo, la variable g podría ser definida de la siguiente manera:

g = 1 si el sujeto es miembro del grupo 1 (grupo focal)

g = 0 si el sujeto es miembro del grupo 2 (grupo de referencia),

El término θg es el producto de las dos variables independientes (nivel de aptitud y grupo). Y finalmente, τ_1, τ_2, y τ_3 son los coeficientes del modelo. El coeficiente τ_1 representa las diferencias de ejecución debidas a la aptitud. Normalmente este coeficiente es significativo, lo cual no es de extrañar ya que es lógico que los sujetos con niveles de aptitud más altos respondan mejor al ítem. El coeficiente τ_2 representa la diferencia de ejecución en el ítem entre los dos grupos, y su interpretación es similar a β_{MH}. Finalmente, el coeficiente τ_3 representa la interacción entre grupo y nivel en el rasgo o aptitud. Estos coeficientes

podrían expresarse según la nomenclatura en la que hemos formulado el modelo anteriormente:

$$\tau_2 = \beta_{01} - \beta_{02}$$

$$\tau_3 = \beta_{11} - \beta_{12}$$

Según esto, un ítem presenta FDI uniforme cuando $\tau_2 \neq 0$, y $\tau_3 = 0$. Es decir, cuando la ejecución en el ítem es diferente en los dos grupos, pero no hay interacción entre el grupo y el nivel de aptitud (las dos curvas tienen la misma pendiente, por lo tanto son paralelas). Por otro lado, un ítem presenta FDI no uniforme cuando $\tau_3 \neq 0$ (independientemente de que $\tau_2 = 0$ o no), ya que esto indica que las dos curvas no son paralelas (no tienen la misma pendiente), por lo tanto hay interacción entre el grupo y el nivel de aptitud.

Swaminathan y Rogers (1990) indican que para confirmar la no existencia de FDI es necesario comprobar que $\tau_2=0$ y que $\tau_3=0$. Estas dos hipótesis pueden ser evaluadas de forma simultánea:

$$\tau_2 = \tau_3 = 0$$

Estos autores señalan que el estadístico para evaluar la hipótesis conjunta presenta una distribución χ^2 con 2 grados de libertad. Cuando el valor de este estadístico excede $\chi^2_{\alpha;gl=2}$, se rechaza la hipótesis de la no existencia de FDI.

Camilli y Shepard (1994) presentan otro procedimiento para el contraste de hipótesis. Estos autores plantean la comparación de

diferentes modelos que no tienen el mismo número de parámetros. El modelo I representa una situación en la que no hay interacción y tampoco hay FDI, por lo tanto $\tau_2 = 0$ y $\tau_3 = 0$. Este es el modelo más sencillo.

Modelo I (no FDI): $$z = \tau_0 + \tau_1\theta$$

El modelo II representa una situación en la que hay FDI uniforme, por lo tanto el coeficiente que representa la interacción entre grupo y nivel de aptitud es igual a cero ($\tau_3 = 0$).

Modelo II (FDI uniforme): $$z = \tau_0 + \tau_1\theta + \tau_2 g$$

Por último, el modelo III que es el más complejo, representa una situación en la que hay FDI no uniforme.

Modelo III (FDI no uniforme): $$z = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g)$$

Estos tres modelos sirven como base para la prueba de hipótesis. Son modelos anidados en los que se añade un nuevo parámetro con respecto al modelo anterior, por lo que el valor χ^2 que indica el ajuste de cada modelo puede ser comparado con el valor χ^2 del modelo en el que se encuentra anidado. La diferencia en los valores de χ^2 para los dos modelos anidados se distribuye con una χ^2 con 1 grado de libertad (igual a la diferencia en parámetros entre los dos modelos). El objetivo es elegir el modelo que se ajuste a los datos y que tenga el menor número de parámetros, es decir, se opera bajo el principio de la parsimonia.

La comparación de los modelos tiene lugar en varios pasos. En primer lugar se examina el ajuste del modelo más sencillo (modelo I). Si este ajusta bien a los datos se procede a evaluar el ajuste del modelo II. Se compara el ajuste de los dos modelos para ver si el modelo II mejora el ajuste ofrecido por el modelo I, y si dicha diferencia es estadísticamente significativa. Si no hay mejora en el ajuste, se puede concluir que el modelo I es el más adecuado, y que por lo tanto no hay FDI. Por otro lado si el modelo II presenta una mejora significativa en el ajuste se pasa a evaluar el ajuste del modelo más complejo (modelo III) y se compara su ajuste con el del modelo II. Si no hay mejora en el ajuste se concluye que el modelo II es más adecuado y esto indica la presencia de FDI uniforme. Si por el contrario el modelo III presenta una mejora significativa en el ajuste se puede concluir la presencia de FDI no uniforme. En la figura 3.9. aparece representado de forma esquemática este proceso de comparación de modelos.

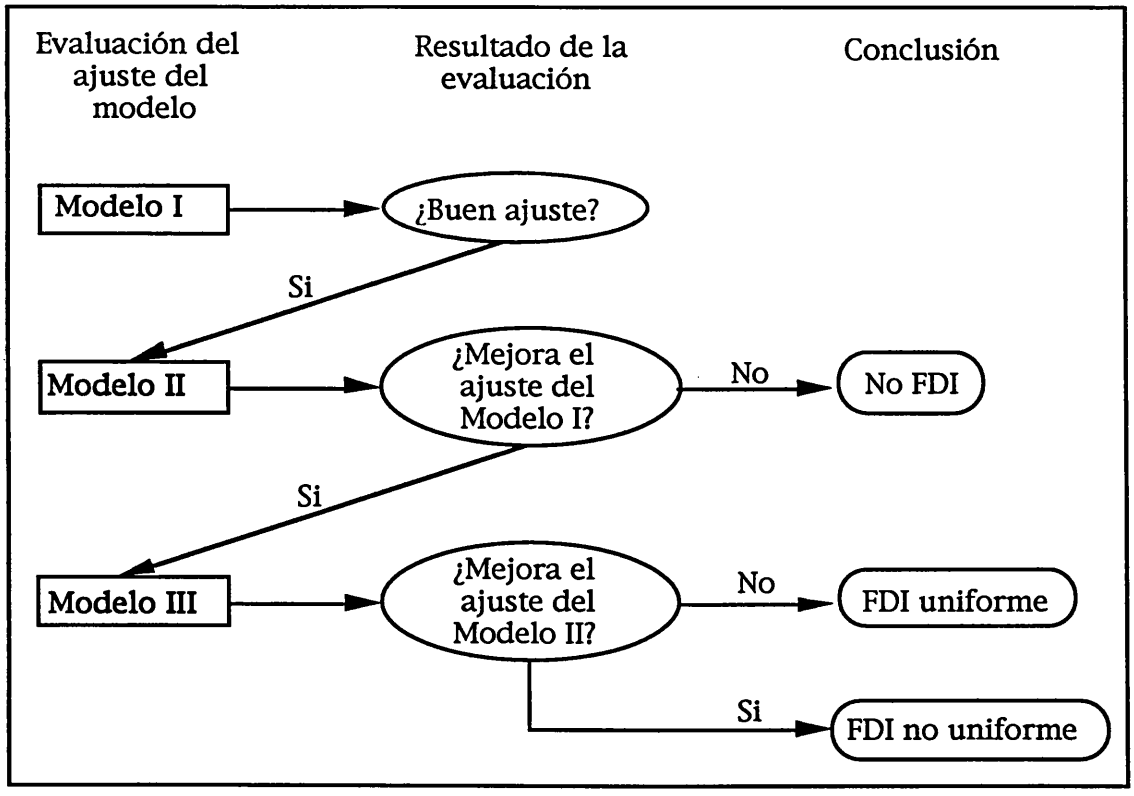


Fig. 3.9. Representación esquemática del proceso de comparación de modelos para el estudio del FDI en la regresión logística.

Los resultados obtenidos con la regresión logística son similares a los obtenidos con el procedimiento Mantel-Haenszel. De hecho cuando la interacción grupo por nivel de aptitud no es incluida en el modelo, el coeficiente τ_2 es similar a β_{MH}. En este caso, ambos son medidas de FDI uniforme.

Una de las ventajas de la regresión logística es que no requiere tamaños muestrales excesivamente grandes. Para aplicar de forma adecuada la regresión logística se requiere que el tamaño de las muestras sea como mínimo de 200 sujetos para cada población. Una de las desventajas que ya ha sido comentada en el apartado anterior es que requiere mucho más tiempo de cálculo que otros procedimientos (como por ejemplo el procedimiento Mantel-Haenszel) ya que los modelos son estimados de forma iterativa.

Análisis Factorial Confirmatorio

A pesar de que algunos de los manuales que recogen los métodos de DFDI no incluyen el análisis factorial entre ellos (Camilli y Shepard, 1994; Holland y Wainer, 1993), son varios los autores que señalan el AFC como un método adecuado para detectar items que presentan funcionamiento diferencial (Millsap y Everson, 1993; Oort, 1992,1996; Gómez Benito, 1996). Su importancia y utilidad es todavía más patente si se tiene en cuenta que todos los métodos de DFDI descritos hasta el momento se basan en el supuesto de la unidimensionalidad del rasgo latente medido por el test. ¿Qué sucede entonces cuando el rasgo latente es multidimensional? ¿Qué métodos se pueden aplicar para detectar el FDI en esos casos? Tanto si se trabaja con medidas continuas como si se trabaja con medidas discretas, el AFC es un método adecuado para

analizar el FDI cuando el rasgo latente es multidimensional. En el caso de trabajar con medidas discretas, los modelos multidimensionales de TRI son otra alternativa adecuada para el estudio del FDI (Millsap y Everson, 1993).

El estudio del FDI mediante AFC puede realizarse desde dos enfoques diferentes: el enfoque de la multidimensionalidad, y el enfoque de la invarianza factorial, también denominados respectivamente contraste del modelo nulo y contraste multimuestra (Gómez Benito, 1996). Oort (1992, 1996) propone el AFC como un método de DFDI basándose en la multidimensionalidad como causa del funcionamiento diferencial de los items en los grupos comparados. Como ya hemos comentado antes, la multidimensionalidad puede explicar que el ítem presente una dificultad diferente en dos grupos que tienen el mismo nivel en la habilidad primaria o constructo medido por el test, pero que presentan una distribución diferente en una habilidad o factor secundarios. Oort analiza el patrón de relaciones entre los items del test, el rasgo latente medido por el test, y los rasgos o factores secundarios que pueden actuar como posibles violadores de la equivalencia a los que denomina violadores potenciales. En la figura 3.10. se presenta un ejemplo que muestra gráficamente estas relaciones. En ella aparece representado un constructo o rasgo latente (T) medido por un test que está formado por p items (x_1-x_p), y dos factores secundarios (V_1 y V_2) que actúan como posibles violadores.

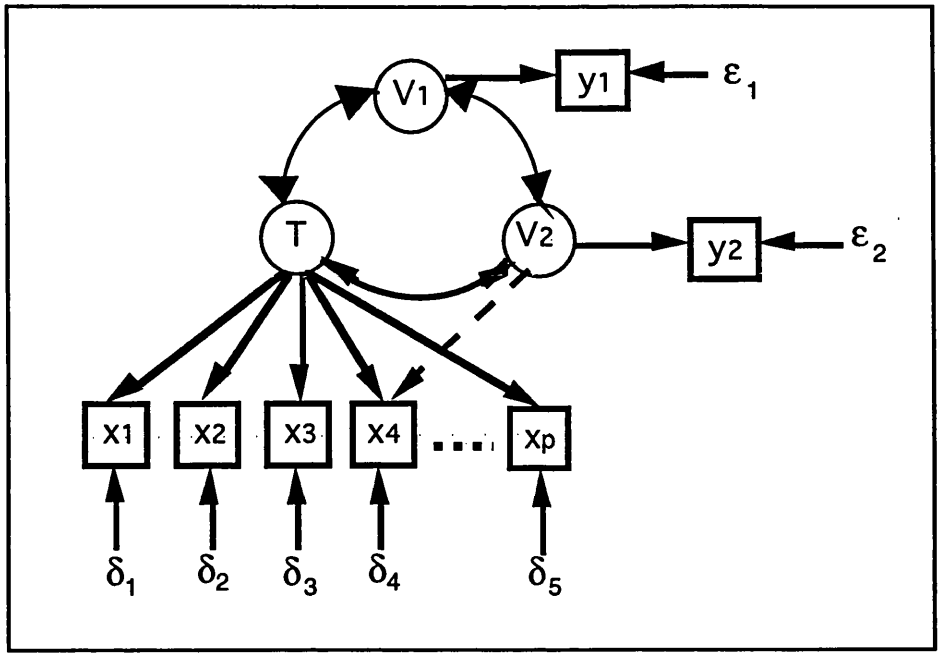


Fig. 3.10. Representación gráfica de las asociaciones entre los items, los rasgos latentes, y los violadores potenciales.

Desde este enfoque el procedimiento que se sigue es el siguiente: se ponen a prueba dos modelos, un modelo base en el que no existe relación entre el ítem y el violador potencial (la saturación factorial se fija a cero), y otro modelo en el que se estima la saturación factorial del ítem en el violador potencial. Si este segundo modelo presenta un ajuste significativamente mejor al del modelo base, se concluye que el ítem presenta FDI. No es necesario evaluar el ajuste de tantos modelos como items tiene el test. El programa LISREL (Jöreskog y Sörbom, 1993) proporciona un índice de modificación para cada ítem que indica cuál sería el incremento del ajuste del modelo respecto al modelo base si la saturación factorial de un ítem i en el violador potencial k fuera liberada para poder ser estimada. El índice de modificación se distribuye aproximadamente con una χ^2 con un grado de libertad. Si el índice de modificación que presenta el valor mayor es significativo, el ítem al que corresponde dicho índice de modificación es eliminado del test. Se vuelve a repetir el proceso de evaluación del ajuste de los dos modelos y

de análisis de los índices de modificación con los items restantes. Este proceso se continúa hasta que los valores de todos los índices de modificación asociados a los items que permanecen sin eliminar son no significativos, lo cual indica que ninguno de éstos items presenta FDI. Este es un procedimiento paso a paso. También es posible eliminar directamente y de una sóla vez todos los items cuyo índice de modificación presenta un valor significativo. Sin embargo, este procedimiento no es muy aconsejable, ya que si el número de items que presentan FDI es muy elevado ello puede llevar por una parte a la no detección de items que presentan FDI, y por otra parte pueden ser eliminados items que realmente no estaban sesgados.

La flecha de trazo discontinuo entre el violador potencial V_2 y el ítem 4 (ver figura 3.10.) indica que existe un efecto directo de dicho violador sobre el ítem. Este efecto directo implica que el ítem 4 no solamente mide el rasgo latente T, sino que también mide V_2. Este ítem presenta FDI respecto al factor secundario V_2, por lo tanto es posible que sujetos con igual nivel en el constructo medido alcancen puntuaciones diferentes en el ítem 4 debido a que sus niveles en V_2 son diferentes.

Como ejemplo para ilustrar este uso del análisis factorial, Oort (1992) evaluó el efecto de cinco violadores potenciales (género, edad, deseabilidad social, motivación y auto-confianza) sobre la puntuación de un grupo de estudiantes en el Cuestionario de Actitudes Escolares de Vorst (1985, 1989) que mide la motivación hacia las tareas escolares, la satisfacción con la vida escolar, y la auto-confianza sobre las propias capacidades escolares.

En el contexto de la comparación transcultural se pone a prueba un modelo base con un rasgo latente (el constructo medido por el test) y un único posible violador (la pertenencia a un grupo cultural determinado), en el que se fijan a cero las saturaciones de los items en el factor "grupo de pertenencia". Para comprobar si los items presentan FDI se formula un modelo alternativo idéntico al anterior pero en el que las saturaciones de los items que presentan un índice de modificación significativo se dejan libres en el factor "grupo de pertenencia" para poder ser estimadas.

El uso del AFC desde la perspectiva de la multidimensionalidad se basa en el análisis de una sola muestra de sujetos, y lo que interesa es analizar la relación entre los items y los posibles violadores. Esto puede tener sus limitaciones, ya que en primer lugar es necesario identificar previamente qué factores pueden actuar como posibles violadores para que sean incluidos en el modelo. Se necesita contar con modelos teóricos fuertes o con estudios empíricos previos que justifiquen la inclusión de esas variables como potenciales violadores; y por otro lado se corre el riesgo de no incluir violadores importantes.

La aportación de Oort es muy reciente, lo cual hace que el uso del AFC como método de DFDI en base a la multidimensionalidad no esté muy extendido. Sin embargo, el estudio de la invarianza factorial entre grupos se utilizó ya a principios de los años 70 (Jöreskog, 1971; McGaw y Jöreskog, 1971), y posteriormente el AFC ha sido frecuentemente utilizado para analizar la invarianza factorial entre diferentes grupos en función del género o la edad (Marsh, 1987a, 1994; Marsh y Hocevar, 1985), o entre grupos culturales diferentes (Reise, Widaman y Pugh, 1993). Desde esta perspectiva se parte de la idea de que las mediciones

psicológicas están en la misma escala, y por lo tanto son comparables, cuando las relaciones empíricas entre el constructo o rasgo medido y los indicadores que lo representan (los items del test) son invariantes entre los grupos (Drasgow y Kanfer, 1985). De este modo, el interés radica en estudiar las relaciones entre los items y el rasgo latente medido por el test, imponiendo diferentes constricciones a ese patrón de relaciones. Los análisis se realizan en base a la diferenciación de dos muestras (que en el contexto de comparación entre culturas representan los grupos culturales a comparar), y se analiza el ajuste de diferentes modelos que implican mayor o menor nivel de constricción de los parámetros de la solución factorial.

Ya hemos visto cómo el AFC es un método adecuado para analizar la equivalencia estructural de diferentes versiones del mismo instrumento. Este método permite poner a prueba la hipótesis de la existencia del mismo patrón de relaciones entre las variables observables (los items) y las variables latentes en los datos procedentes de diferentes grupos culturales. Si se confirma dicha hipótesis, ello indicaría que subyacen las mismas variables latentes en los grupos comparados. Este sería un primer paso en el estudio de la equivalencia, y una garantía necesaria, aunque no suficiente, para asegurar la validez de las comparaciones entre grupos.

En este primer paso se pone a prueba la equivalencia del patrón de relaciones, pero no se impone ninguna constricción a los parámetros. El uso del AFC como método de DFDI supone ir un poco más lejos y añadir una serie de constricciones a ese patrón de relaciones: diferentes parámetros pueden ser constreñidos de tal forma que sean invariantes (es decir, equivalentes) en los grupos comparados, y de este modo poder

evaluar el ajuste diferencial de modelos anidados en los que se van incluyendo progresivamente mayor número de constricciones. Esto permite poner a prueba un conjunto de hipótesis sobre la invarianza o equivalencia de determinados componentes de la solución factorial. Las hipótesis a evaluar, con diferentes grados de constricción, podrían ser la siguientes:

1° Invarianza de la estructura factorial (equivalencia estructural).

2° Invarianza de las saturaciones factoriales.

3° Invarianza de las saturaciones factoriales, de las varianzas de los factores y de las covarianzas entre factores.

4° Invarianza de las saturaciones factoriales, de las varianzas y covarianzas entre factores, y de los términos de error.

Programas como LISREL (Jöreskog y Sörbom, 1993) permiten calcular diferentes índices de ajuste que evalúan si la matriz de covarianzas reproducida a partir de los parámetros estimados $(\hat{\Sigma})$, difiere significativamente de la matriz de covarianzas observada en la muestra (S). Es importante señalar que este tipo de análisis puede realizarse utilizando tanto la matriz de covarianzas como la matriz de correlaciones de las variables observadas. Sin embargo, en el contexto de análisis multigrupo lo adecuado es trabajar con la matriz de covarianzas, y no con la de correlaciones. La comparación de parámetros entre grupos solamente es posible si las variables presentan una métrica común en todos los grupos. Si se utilizara la matriz de correlaciones, ello supondría que se han estandarizado los datos separadamente para cada grupo, y por lo tanto las variables no presentarían una escala común entre los grupos. En tal caso, la comparación de los parámetros entre grupos no sería posible.

En el contexto transcultural se llevan a cabo análisis multigrupo que permiten comparar el ajuste de diferentes modelos. No hay un consenso claro sobre el orden a seguir en la evaluación de los modelos (Marsh, 1994). Diferentes autores recomiendan evaluar la invarianza de los términos de error antes que la invarianza de las varianzas y covarianzas entre factores (Jöreskog, 1971; Jöreskog y Sörbom, 1988). Otros, aún haciendo esta misma recomendación, plantean que el orden es realmente arbitrario (Bollen, 1989). Sin embargo, otros autores indican que la hipótesis menos relevante es la que evalúa la equivalencia de los parámetros de los términos de error, y por lo tanto recomiendan evaluar la invarianza de las varianzas y covarianzas entre factores antes que la invarianza de los términos de error (Bentler, 1988; Byrne, 1989; Marsh, 1987a, 1994).

A pesar de esta falta de consenso, lo que sí parece claro es que la condición mínima para establecer la invarianza factorial es la equivalencia de las saturaciones factoriales. También parece más adecuado establecer una jerarquía de modelos anidados que facilite la comparación del ajuste de los diferentes modelos. Si el análisis de estos modelos se lleva a cabo de forma jerárquica, cada uno de ellos se encuentra anidado en el modelo anterior. La diferencia en los valores de χ^2 para dos modelos anidados se distribuye con una χ^2 con grados de libertad igual a la diferencia en grados de libertad de los dos modelos. Por lo tanto, el ajuste de cada modelo se puede comparar con respecto al modelo anterior en el cual se encuentra anidado. En base a las hipótesis formuladas anteriormente, y siguiendo este planteamiento de modelos jerárquicos anidados, los modelos a analizar serían los siguientes:

1° Modelo 1 para poner a prueba la hipótesis de la invarianza de la estructura factorial: la comprobación de esta hipótesis supone comprobar el ajuste de un modelo donde se evalúa si las dos muestras presentan el mismo número de variables latentes con las mismas especificaciones sobre parámetros libres o fijos en la matriz lambda X de saturaciones factoriales (Λ_x), en la matriz phi de varianzas y covarianzas entre factores (Φ), y en la matriz theta-delta de términos de error (Θ_δ). En este modelo únicamente se especifica si los datos de los dos grupos culturales presentan el mismo número de factores o rasgos latentes y si el patrón de relaciones entre los indicadores del rasgo (ítems) y los rasgos latentes medidos por el test es idéntico en ambos grupos. Este modelo es el menos restrictivo y representa el modelo base a partir del cual se evalúa el ajuste de los sucesivos modelos que incluyen mayor número de constricciones.

2° Modelo 2 para poner a prueba la hipótesis de la invarianza de las saturaciones factoriales: este sería un modelo más restrictivo que el anterior en el que se constriñen las saturaciones factoriales de la matriz lambda X (Λ_x) a ser invariantes a través de los diferentes grupos comparados. Este modelo se encuentra anidado en el modelo base, por lo que el valor de χ^2 que indica su ajuste se compara con el valor de χ^2 del modelo 1 o modelo base. Si la diferencia entre las χ^2 de los modelos 1 y 2 no es estadísticamente significativa, ello confirmaría la hipótesis de la invarianza de las saturaciones factoriales entre los grupos comparados. Es decir, si se comparan g grupos, se confirmaría que $\Lambda_x^{(1)} = \Lambda_x^{(2)} = \ldots = \Lambda_x^{(g)}$.

3° Modelo 3 para poner a prueba la hipótesis de la invarianza de las saturaciones factoriales, de las varianzas de los factores y de las

covarianzas entre factores: este modelo recoge la constricción impuesta en el modelo 2 respecto a la invarianza de las saturaciones factoriales, y además constriñe los parámetros de la matriz de varianzas y covarianzas entre factores (Φ) a ser invariantes a través de los grupos. Por tanto, la hipótesis que se pone a prueba es la siguiente:

$$\Lambda_x^{(1)} = \Lambda_x^{(2)} = \ldots = \Lambda_x^{(g)} \quad \text{y} \quad \Phi^{(1)} = \Phi^{(2)} = \ldots = \Phi^{(g)}$$

Si la diferencia entre la χ^2 de este modelo y la del modelo 2 (en el cual se encuentra anidado) no es estadísticamente significativa, ello confirmaría la hipótesis que se pone a prueba.

4° Modelo 4 para poner a prueba la hipótesis de la invarianza de las saturaciones factoriales, de las varianzas y covarianzas entre factores, y de los términos de error: este es el modelo más restrictivo, ya que recoge las constricciones impuestas en el modelo 3, y además se constriñen los parámetros de la matriz theta-delta de términos de error (Θ_δ) a ser invariantes entre los grupos. Normalmente se supone que los errores de medida de las variables observadas no están correlacionados, por lo que la matriz theta-delta es una matriz cuyos elementos no diagonales son igualados a cero. Poner a prueba este modelo supone fijar las saturaciones factoriales de la matriz lambda X (Λ_x), los parámetros de la matriz de varianzas y covarianzas entre factores (Φ), y los parámetros de la matriz theta-delta de términos de error (Θ_δ) a ser invariantes a través de los grupos. En este caso, la hipótesis a evaluar plantea que:

$$\Lambda_x^{(1)} = \Lambda_x^{(2)} = \ldots = \Lambda_x^{(g)}, \quad \Phi^{(1)} = \Phi^{(2)} = \ldots = \Phi^{(g)}, \quad \text{y} \quad \Theta_\delta^{(1)} = \Theta_\delta^{(2)} = \ldots = \Theta_\delta^{(g)}$$

El contraste de la hipótesis se establece evaluando la significación estadística de la diferencia entre las χ^2 del modelo 4 y del modelo 3 en el cual se encuentra anidado.

Esta jerarquía de modelos e hipótesis (con el mismo o diferente orden) que plantean diferentes autores (Jöreskog, 1971; Jöreskog y Sörbom, 1988; Bollen, 1989; Bentler, 1988; Byrne, 1989; Marsh, 1987a, 1994; Marsh y Hocevar, 1985), no resulta necesaria desde la perspectiva de otros. Por ejemplo, Reise et al. (1993), plantean que para probar la invarianza de las medidas entre grupos, y por lo tanto que las puntuaciones de ambos grupos son comparables, basta con comprobar la invarianza de las saturaciones factoriales. Estos autores no hacen restricciones sobre las varianzas y covarianzas de las variables latentes, porque según indican, es probable que los grupos difieran con respecto a ellas. MacCallum y Tucker (1991) argumentan que en principio las saturaciones factoriales deben ser invariantes entre los grupos, pero las varianzas y covarianzas factoriales pueden tener cierto grado de especificidad dentro de cada muestra. Desde este enfoque se plantea también el ajuste de diferentes modelos pero el nivel de restricción se establece siempre en base a la invarianza de las saturaciones factoriales. Las hipótesis a contrastar serían las siguientes:

1° Invarianza de la estructura factorial (equivalencia estructural).
2° Invarianza total de las saturaciones factoriales.
3° Invarianza parcial de las saturaciones factoriales.

Al igual que en el planteamiento anterior, desde este enfoque se llevan a cabo análisis multigrupo que permiten comparar el ajuste de

diferentes modelos para poner a prueba las hipótesis formuladas anteriormente. Los modelos implicados son los siguientes:

1° Modelo 1 para poner a prueba la hipótesis de la invarianza de la estructura factorial: el modelo analizado permite estudiar la equivalencia estructural de los datos. Los valores de todos los parámetros se estiman libremente, y este modelo sirve como modelo base, ya que su ajuste sirve como referencia para comparar el ajuste de los modelos más restrictivos que están anidados en él.

2° Modelo 2 para poner a prueba la hipótesis de la invarianza total de las saturaciones factoriales: se pone a prueba el ajuste de un modelo que plantea la invarianza de todas las saturaciones factoriales de la matriz lambda X (Λ_x). Es decir, si estamos comparando g grupos culturales, la hipótesis a confirmar plantea la invarianza de las matrices lambda para todos los grupos:

$$\text{Ho: } \Lambda_x^{(1)} = \Lambda_x^{(2)} = \ldots = \Lambda_x^{(g)}$$

Se compara el valor de χ^2 de este modelo con el valor de χ^2 obtenido en el modelo base. La diferencia en los valores de χ^2 para los dos modelos anidados se distribuye con una χ^2 con grados de libertad igual a la diferencia en grados de libertad de los dos modelos. Si el modelo 2 no introduce un incremento significativo en χ^2 con respecto al modelo base se confirma la hipótesis de invarianza total de las saturaciones factoriales. Si por el contrario el modelo 2 introduce un cambio significativo en χ^2, ello indicaría que todos los items no se relacionan con el constructo de la misma manera en los diferentes grupos comparados. Se debe rechazar la hipótesis de invarianza total, y

se pasa a analizar una 3ª hipótesis que plantea la invarianza parcial de las saturaciones factoriales.

3° Modelo 3 para poner a prueba la hipótesis de la invarianza parcial de las saturaciones factoriales: si se rechaza la hipótesis de la invarianza total, se puede pasar a evaluar si algún subconjunto de items son invariantes. La identificación de este subconjunto de items invariantes se realiza en base a los índices de modificación que se calculan para cada parámetro fijado o restringido. Los valores de estos índices indican cuánto cambiaría el valor de χ^2 asociado al modelo si se eliminara la restricción de invarianza sobre ese parámetro. Si el índice de modificación de una saturación factorial determinada es significativo, eso indica que la saturación de este ítem no es invariante entre los diferentes grupos culturales. Por lo tanto, poner a prueba la hipótesis de invarianza parcial implica analizar el ajuste de un modelo con las mismas especificaciones que el modelo de invarianza total pero sin imponer restricciones de invarianza sobre las saturaciones factoriales que presentan índices de modificación significativos. Los valores de ajuste de este modelo se comparan con el modelo base en el que se encuentra anidado. Si el modelo de invarianza parcial no difiera significativamente del modelo base se puede concluir que es más adecuado el primer modelo ya que es más parsimonioso.

Por otro lado, no es necesario confirma la hipótesis de invarianza total para concluir que las puntuaciones de los diferentes grupos culturales son comparables. La comparación de los grupos es posible también si se confirma la hipótesis de la invarianza parcial (Byrne, Shavelson y Muthén, 1989), es decir, si algunas, aunque no todas las saturaciones factoriales son invariantes entre los grupos.

Reise et al. (1993) comparan el AFC y los procedimientos basados en la TRI como métodos para analizar el funcionamiento diferencial de los items. Estos autores señalan que las diferencias encontradas entre estos dos procedimientos son debidas entre otras cosas, a que el AFC analiza la invarianza de la saturación factorial o parámetro lambda que correspondería al parámetro de discriminación del ítem, mientras que la TRI analiza conjuntamente la invarianza del parámetro de discriminación y del parámetro de dificultad del ítem. Desde esta perspectiva se puede concluir que los procedimientos basados en la TRI imponen más restricciones de invarianza en el estudio del FDI, pero tal y como Millsap y Everson (1993) señalan, la invarianza de los parámetros de dificultad también puede ser analizada mediante el AFC. Para ello sería necesario realizar el análisis de invarianza utilizando un modelo de estructura de medias latentes. Vamos a pasar a comentar el uso del AFC desde este enfoque, al mismo tiempo que justificamos la similitud establecida entre los parámetros de la TRI y los del AFC.

El AFC asume que la relación entre las variables observables y las variables latentes es lineal. Esto se refleja en el modelo general que asume el AFC y que ya ha sido comentado en páginas anteriores:

$$X = \Lambda_x \xi + \delta$$

Este modelo asume que tanto las medias de las variables observables como las medias de las variables latentes presentan un valor igual a cero. El análisis se realiza en base a la matriz de covarianzas o de correlaciones de las variables observables, por lo que para evaluar el ajuste del modelo a los datos, únicamente se tiene en cuenta la estructura de covarianzas, pero no la estructura de medias. Sin

embargo, se puede formular un modelo que incorpore los valores de las medias de las variables observables y de las variables latentes (Jöreskog y Sörbom, 1985; Sörbom, 1974, 1978; 1982):

$$X = \tau_x + \Lambda_x \xi + \delta$$

donde,

- X es un vector de orden (q x 1) compuesto por q variables observables,
- τ_x es un vector de orden (q x 1) de parámetros interceptos,
- Λ_x es una matriz de orden (q x r) compuesta por las saturaciones factoriales que expresan la relación de cada variable observable con la correspondiente variable latente,
- ξ es un vector de orden (r x 1) compuesto por r variables latentes, y
- δ es un vector de orden (q x 1) que representa el error de medición de cada variable observable.

Se puede expresar el valor esperado de las variables observables en función de las medias de las variables latentes:

$$E(X) = \tau_x + \Lambda_x \kappa \qquad (3.1)$$

donde κ es un vector de orden (r x 1) que contiene las medias de las variables latentes $(E(\xi) = \kappa)$.

Para estimar este modelo es necesario introducir una serie de modificaciones con respecto al modelo general. Como se puede observar, el modelo de AFC con medias latentes introduce dos nuevos vectores de

parámetros: el vector tau-x de parámetros interceptos (τ_x), y el vector kappa de medias de las variables latentes (κ). También es necesario incluir las medias de las variables observables en la matriz de entrada de datos. Los parámetros a estimar con el programa LISREL en un análisis multigrupo son los siguientes:

* Lambda-x $\lambda_x^{(1)}, \lambda_x^{(2)}, \ldots, \lambda_x^{(g)}$, matrices de saturaciones factoriales.

* Tau-x $\tau_x^{(1)}, \tau_x^{(2)}, \ldots, \tau_x^{(g)}$, vectores de parámetros interceptos.

* Kappa $\kappa^{(1)}, \kappa^{(2)}, \ldots, \kappa^{(g)}$, vectores de medias de las variables latentes (ξ).

* Phi $\Phi^{(1)}, \Phi^{(2)}, \ldots, \Phi^{(g)}$, matrices de covarianzas entre variables latentes (ξ).

* Theta-delta $\Theta_\delta^{(1)}, \Theta_\delta^{(2)}, \ldots, \Theta_\delta^{(g)}$, matrices de covarianzas entre los errores.

Desde este enfoque, el estudio del funcionamiento diferencial de los items implica analizar la invarianza de las matrices de saturaciones factoriales y de los vectores de parámetros interceptos entre los grupos a comparar. Es decir, la hipótesis a evaluar plantea que:

$$\Lambda_x^{(1)} = \Lambda_x^{(2)} = \ldots = \Lambda_x^{(g)} \quad \text{y} \quad \tau_x^{(1)} = \tau_x^{(2)} = \ldots = \tau_x^{(g)}$$

Tenemos todavía pendiente establecer la similitud establecida entre los parámetros de la TRI y del AFC, y por lo tanto la aportación que hace el AFC con medias latentes al estudio del FDI. Según la ecuación

(3.1.), el intercepto (τ_x) representa el valor esperado de la variable observable cuando κ vale cero. Es decir, representa el valor esperado de respuesta en el ítem para aquellos sujetos que tienen un valor de 0 en el rasgo latente, por lo tanto puede ser interpretado como el parámetro de dificultad del ítem; por otra parte, la saturación factorial representa la pendiente de la línea de regresión, y puede ser interpretada como el parámetro de discriminación del ítem (Mellenbergh, 1994). Vamos a ver esto más claramente con un ejemplo gráfico.

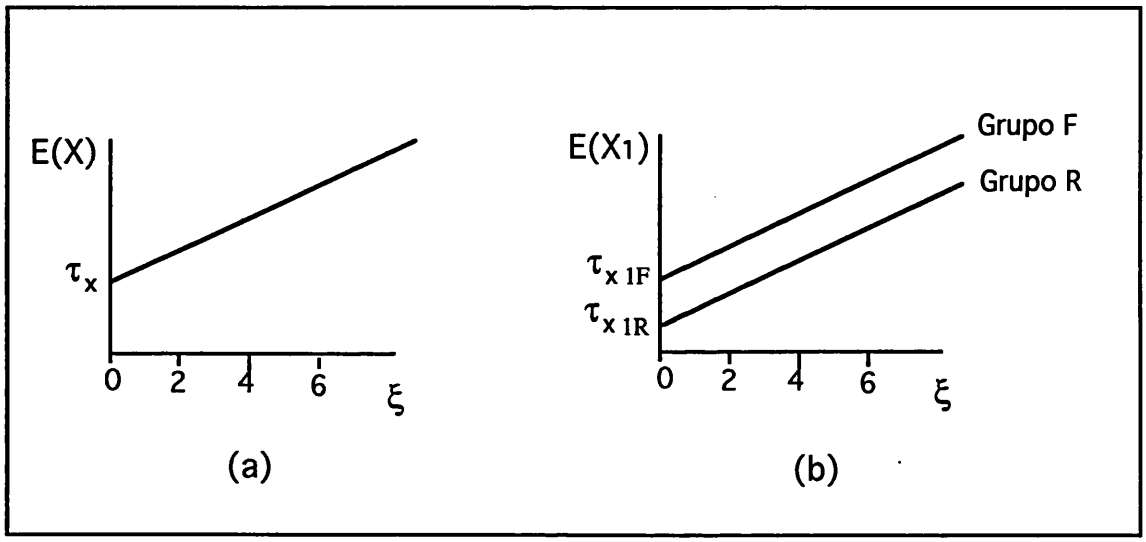


Fig. 3.11. Representación gráfica del intercepto.

En la figura 3.11.a se observa cómo el intercepto (τ_x) representa el valor esperado de la variable observable E(X) para aquellos sujetos que presentan un valor de 0 en el rasgo latente. Por otro lado, la gráfica 3.11.b representa la línea de regresión del valor esperado en el ítem X_1 para dos grupos diferentes. En este ejemplo λ_x es invariante, es decir, la pendiente de la línea de regresión para los grupos focal y de referencia es la misma. Sin embargo, el intercepto es diferente: el valor esperado de la variable observable para los sujetos que tienen un valor de 0 en el rasgo latente es mayor en el grupo focal, o lo que es lo mismo, para un

mismo nivel de ξ, el ítem resulta más dificil para los sujetos del grupo de referencia

Esta interpretación del intercepto como el parámetro de dificultad del ítem se hace en el contexto de un modelo de respuesta continua al ítem (Mellenbergh, 1994; Ferrando, 1996b). Sin embargo, se puede considerar que la escala de respuesta continua representa un caso extremo de la escala de respuesta graduada con un elevado número de opciones de respuesta. Por lo tanto, esta interpretación sobre el intercepto es extrapolable a escalas de respuesta graduada, como la escala tipo Likert (Mellenberg, 1994). Los trabajos de Thissen, Steinberg, Pyszczynski, y Greenberg (1983), Millsap y Everson (1991), y Everson, Millsap, y Rodriguez (1991), representan algunos ejemplos del uso del AFC con medias latentes utilizando escalas de respuesta tipo Likert.

La aportación del AFC con medias latentes al estudio del FDI es evidente. Si las saturaciones factoriales y/o los interceptos difieren entre grupos, a sujetos que presentan un mismo nivel de puntuación en el rasgo latente les corresponderán diferentes valores esperados de la variable observable en función del grupo al que pertenezcan. El modelo de estructura de covarianzas sólo considera el parámetro de discriminación del ítem, es decir, únicamente permite analizar la invarianza de las saturaciones factoriales. Sin embargo, el modelo de medias latentes permite poner a prueba la invarianza de las saturaciones factoriales y de los interceptos en diferentes grupos. Si la línea de regresión presenta pendientes diferentes en los dos grupos, es decir, si no se confirma la invarianza de las saturaciones factoriales en los dos grupos, ello indicaría la presencia de FDI no uniforme; por otro lado, si la línea de regresión difiere solamente en los interceptos pero no

en las pendientes, ello indicaría la presencia de FDI uniforme (Mellenbergh, 1994).

El uso del AFC con medias latentes no es algo nuevo (Sörbom, 1974), sin embargo no son muchos los trabajos en los que ha sido utilizado. Byrne y cols. (1989) tras una revisión de la literatura encontraron únicamente dos trabajos empíricos publicados en los que se analizaban las diferencias entre grupos utilizando la estructura de medias latentes (Lomax, 1985; McGaw y Jöreskog, 1971); posteriormente tampoco se ha difundido en exceso el uso de esta metodología (Byrne et al., 1989; Cole y Maxwell, 1985; Everson et al., 1991; Ferrando, 1996b; Millsap y Everson, 1991; Muthén y Christoffersson, 1981). Sin embargo, todos estos trabajos ponen de manifiesto la utilidad del AFC con medias latentes para analizar el FDI.

El AFC con medias latentes permite poner a prueba la invarianza de las saturaciones factoriales y de los interceptos entre grupos diferentes. También es posible plantear hipótesis sobre la invarianza parcial (Byrne et al., 1989) en las que algunos de los parámetros de los interceptos se mantengan invariantes entre los grupos. El uso de esta metodología permite detectar diferencias entre los grupos que no serían detectables si se analiza únicamente la estructura de covarianzas. Pero para poner a prueba las hipótesis de la invarianza conjunta de las saturaciones factoriales y los interceptos es necesario haber probado previamente la invarianza de las saturaciones factoriales utilizando una estructura de covarianzas. El modelo de invarianza de saturaciones e interceptos está anidado en el modelo de invarianza de las saturaciones, por lo que es posible comparar el ajuste de ambos modelos mediante un test de la diferencia entre χ^2. En el caso de no confirmarse la hipótesis

de la invarianza total, los índices de modificación ofrecen información relevante para plantear hipótesis de invarianza parcial.

La importancia del estudio del funcionamiento diferencial de los items antes de pasar a analizar las diferencias transculturales es evidente. Se debe identificar los items que muestran diferencias, y analizarlos en detalle para determinar las posibles explicaciones de estas diferencias. Una adaptación deficiente del ítem, una traducción inadecuada, o el uso de términos o situaciones que son desconocidas o poco familiares en esa población, podrían ser algunas de las posibles explicaciones. Sin embargo, este tipo de análisis no son habitualmente realizados en la investigación transcultural (Van de Vijver y Leung, 1996). Esto puede ser debido, en parte, a los problemas que surgen de la aplicación de este tipo de técnicas. En primer lugar, puede darse el caso de que se detecte FDI, pero que resulte difícil encontrar la explicación a ese funcionamiento diferencial. De hecho, no es extraño encontrar estudios en los que no se acierta a dar una explicación adecuada al funcionamiento diferencial detectado en ciertos items (Scheuneman, 1987; Van de Vijver, 1994). En segundo lugar, los resultados del análisis de los items no suelen ser estables en los estudios de validación transcultural. Diferentes estudios con el mismo instrumento pueden ofrecer resultados en los que los items que presentan FDI no sean los mismos. Finalmente, algunos estudios sobre sesgo de los items han informado de una alta proporción de items que presentan funcionamiento diferencial, llegando algunas veces a más de la mitad de los items, lo cual indicaría un serio problema de validez del instrumento. Esta falta de estabilidad en los resultados y los problemas de

interpretación que surgen en la aplicación empírica, pueden ser en parte responsables del menor uso de las métodos de DFDI.

CAPITULO 4

AUTOCONCEPTO FISICO

Vamos a dedicar este capítulo a ofrecer una panorámica general sobre la conceptualización y el estudio del autoconcepto, así como de los instrumentos desarrollados para su medida. Nos centraremos principalmente en el autoconcepto físico, ya que es éste el constructo medido por el Cuestionario de Autoconcepto Físico -"Physical Self Description Questionnaire"-, cuestionario que se analiza en el presente trabajo.

1. EL AUTOCONCEPTO

En este primer apartado presentamos las aportaciones de los primeros autores que reflexionaron sobre el autoconcepto, su naturaleza, y su formación: William James, pionero en la conceptualización del self y del autoconcepto; y desde el enfoque del interaccionismo simbólico, Cooley y Mead. A continuación presentamos los diferentes modelos teóricos que han sido propuestos sobre la estructura del autoconcepto, para pasar después a desarrollar el modelo multidimensional y jerárquico propuesto por Shavelson, Hubner y Stanton (1976).

1.1. Primeras aportaciones a la conceptualización del autoconcepto

Si tuviéramos que atribuir a alguien la paternidad del autoconcepto, ésta le correspondería por derecho a William James. Su libro "Principios de Psicología" (James, 1890/1963), considerado como el primer libro de texto introductorio a la psicología (Marsh y Hattie, 1996), dedica el capítulo más extenso a la conceptualización, definición y análisis del autoconcepto. La contribución de James en este trabajo y en trabajos posteriores (James, 1892) hace que se le reconozca como el primer psicólogo que desarrolló una teoría del autoconcepto.

Entre las principales aportaciones de James a la conceptualización del autoconcepto podemos citar: 1) la distinción entre un Yo-self y un Mi-self; 2) una conceptualización multidimensional y jerárquica del Mi-self; y 3) una formulación de las causas de los diferentes niveles de autoestima en los individuos. En primer lugar, este autor hace la distinción entre dos aspectos fundamentales del self, el "Yo" y el "Mi",

que representan respectivamente el self como sujeto (Yo-self), y el self como objeto (Mi-self). Para James, el Yo-self es el conocedor, mientras que el Mi-self es el objeto de conocimiento, y hace referencia al conocimiento que el Yo tiene sobre sí mismo. De este modo, el Yo es el agente activo responsable de la construcción del Mi. De estos dos componentes del self, el que ha recibido mayor atención y ha originado mayor número de estudios y publicaciones es el Mi-self, más conocido actualmente como autoconcepto.

James (1890/1963), también presta especial atención al Mi-self, y la conceptualización que ofrece de él deja ya clara su naturaleza multifacética y jerárquica. Distingue tres constituyentes del Mi-self: el self material, el self social y el self espiritual. El self material incluye el self corporal y la posesiones, es decir, todo aquello que uno puede reconocer como propio ("mi cuerpo", "mis características corporales", "mis pertenencias materiales"). El self social consiste en aquellas características del self reconocidas por los otros, y según James (1890/1963), "una persona tiene tantos selfs sociales como individuos que le reconocen y tienen una imagen de él en su mente" (pag. 190). Por último, el self espiritual es definido como los pensamientos, inclinaciones, juicios de valor, etc., del individuo, todos ellos considerados como los aspectos más estables del self.

Además de dimensionalizar el self, James impuso una estructura jerárquica a sus componentes. Según este autor, en la base de esta jerarquía se encuentra el self material. El self social ocupa una posición intermedia, asumiendo que el individuo da mayor importancia a la imagen que los otros puedan tener de él, que a su propio cuerpo o a sus riquezas. Por último, el self espiritual ocupa la cima de esta jerarquía,

representando su importancia por encima de los otros dos componentes. Como hemos comentado anteriormente, con esta conceptualización James abre el camino para el desarrollo de modelos posteriores en los que el self se conceptualiza como un constructo multidimensional y jerárquico.

Otra importante aportación de James (1890/1963) es su definición de la autoestima como la ratio entre los éxitos y las pretensiones de un sujeto. Según este autor, la autoestima no se puede reducir simplemente a la suma de los éxitos percibidos en la vida. Para su conceptualización y comprensión, es necesario tener en cuenta las aspiraciones de éxito del sujeto, de tal modo que si el individuo percibe que los éxitos obtenidos igualan o superan sus pretensiones o aspiraciones de éxito, esto originará altos niveles de autoestima. Por el contrario, si las pretensiones son mayores que los éxitos alcanzados, es decir, si la persona no tiene éxito en dominios en los que tenía grandes aspiraciones, ello originará bajos niveles de autoestima. Esta última idea pone de manifiesto el papel de la importancia asignada por el sujeto a un área o dominio concreto. La falta de éxito en un área en la que la persona no tiene pretensiones, no afectará negativamente a su autoestima. Esto lo ejemplifica James sobre sí mismo: "Yo, que a lo largo del tiempo he puesto mi mayor empeño en ser un buen psicólogo, me siento humillado si otros saben más psicología que yo. Sin embargo, estoy contento revolcándome en la más absoluta ignorancia del griego" (pag. 310). Otro ejemplo en palabras del mismo autor: "Tenemos la paradoja de un hombre avergonzado y humillado porque solamente es el segundo boxeador o el segundo remero del mundo... Sin embargo, su débil compañero, que puede ser derrotado por cualquiera, no sufre ningún disgusto a causa de ello, porque hace tiempo que abandonó el intento de 'seguir ese camino'" (pag. 310).

Por su parte, el interaccionismo simbólico pone mayor énfasis en el modo en que las interacciones sociales del individuo con los otros configuran su self. Desde esta perspectiva, el self es considerado como una construcción social desarrollada a través de los intercambios lingüísticos (interacciones simbólicas) con los otros (Cooley, 1902; Mead, 1925, 1934). Las reflexiones de Cooley (1902) sobre el autoconcepto, recuerdan el self social descrito por William James, y su mayor contribución se resume en la metáfora del espejo. Para Cooley, los otros significativos constituyen un espejo social que contribuye a la formación del autoconcepto. De este modo, la percepción que uno tiene sobre sí mismo está determinada por su percepción de las opiniones que los otros tienen sobre él. Lo que viene a constituir el self es lo que imaginamos que los otros piensan de nuestra apariencia, actos, intereses, etc. Para Cooley, en el desarrollo del autoconcepto (o "self-idea" como el autor lo denomina) confluyen tres elementos: 1) la idea que el individuo se forma sobre cómo es su apariencia para la otra persona (¿cómo me percibe el otro?), 2) la idea que el individuo se forma acerca de la opinión o valoración de la otra persona sobre su apariencia (¿qué opinión le merece al otro mi apariencia?), y 3) una especie de sentimiento hacia el propio self, como orgullo o humillación (¿qué sentimiento me provoca la valoración que el otro hace de mí?).

Mead (1925, 1934), recoge las aportaciones de Cooley poniendo una mayor insistencia en el rol de la interacción social, particularmente a través del uso del lenguaje. Al igual que James, distingue entre un Yo-self y un Mi-self. Su atención se centra en el Mi-self, al que confiere un carácter primordialmente social. Mead defiende la idea de que el self es el resultado de un proceso social, y que el lenguaje, en forma de gesto vocal, es el que posibilita el mecanismo para su emergencia.

1.2. Modelos sobre la estructura del autoconcepto

A pesar de que las aportaciones de William James y del interaccionismo simbólico representaron abrir una importante vía para el estudio del autoconcepto, éste estuvo prácticamente olvidado por los investigadores hasta la década de los 50. ¿A qué se debía esa falta de interés por el estudio del autoconcepto? La respuesta es fácil si tenemos en cuenta que durante los años 30 y 40 se produce el desarrollo del conductismo. Desde un enfoque conductista, la investigación se centra en aquellos aspectos de la conducta que pueden ser observados y medidos. El autoconcepto no es un constructo directamente observable, es una experiencia interna y una interpretación subjetiva, por tanto, no recibe atención por parte de los investigadores conductistas. Posteriormente, en la década de los 50 se produce una mayor difusión de las primeras conceptualizaciones del autoconcepto, y éste vuelve a ser objeto de interés por los investigadores. El autoconcepto es definido como un constructo cognitivo (Rogers, 1951; Sarbin, 1952) y afectivo (Sullivan, 1953), y empiezan a desarrollarse instrumentos para su medición.

Las diferentes definiciones del autoconcepto, así como los instrumentos elaborados para su medición, han sido desarrollados a partir de modelos teóricos que hipotetizan una estructura determinada del autoconcepto y de los elementos que lo componen. Marsh y Hattie (1996) y Marsh (1997) describen seis modelos diferentes de la estructura del autoconcepto que han sido discutidos en la literatura, y que se basan en modelos análogos desarrollados en la investigación de la inteligencia. Estos modelos aparecen representados en la figura 4.1.

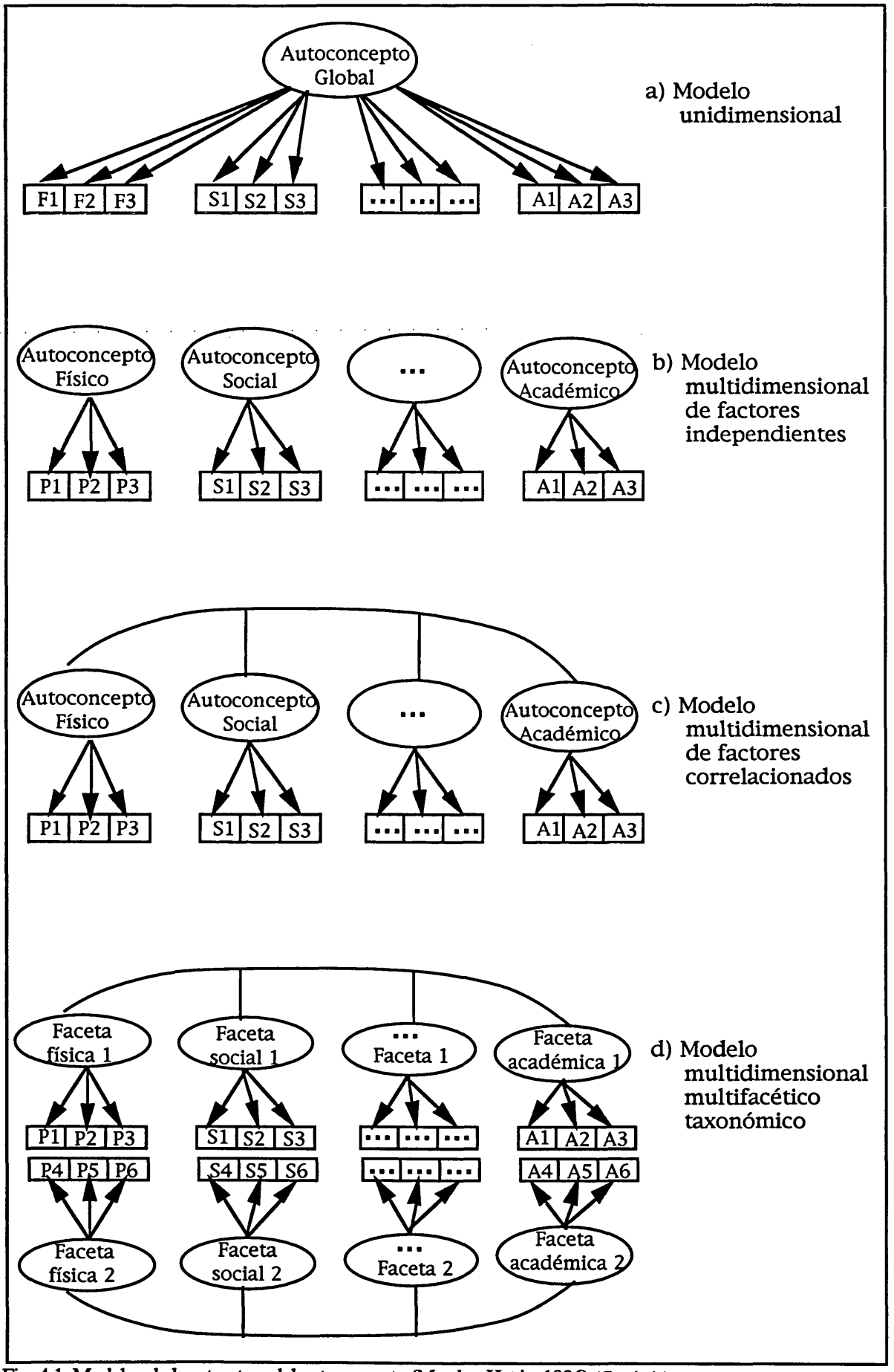


Fig. 4.1. Modelos de la estructura del autoconcepto (Marsh y Hattie, 1996).(Continúa).

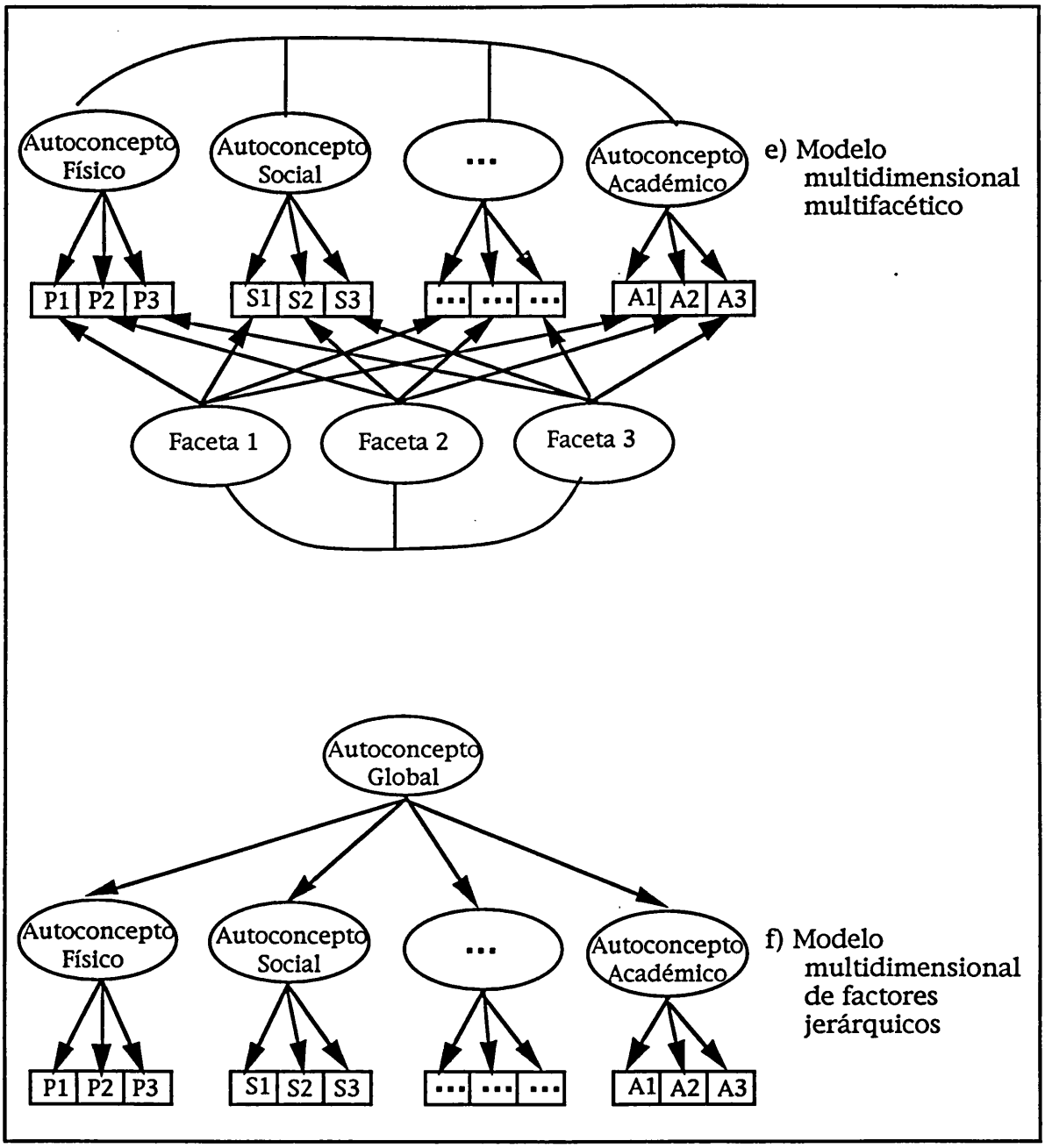


Fig. 4.1. (continuación)

El modelo unidimensional, se asemeja al modelo de Spearman de inteligencia, y sugiere que hay un único factor general de autoconcepto o que un factor general domina sobre otros factores más específicos. Por ejemplo, Coopersmith (1967) y Marx y Winne (1978) argumentaron que las múltiples dimensiones del autoconcepto estaban tan fuertemente dominadas por un factor general que no podían diferenciarse de forma adecuada factores separados. Sin embargo, análisis posteriores de los

estudios llevados a cabo por estos autores ponen de manifiesto que éstos presentan importantes debilidades metodológicas (Dyer, 1964; Marsh y Smith, 1982; Shavelson y Bolus, 1982), y por lo tanto no sirven como evidencia para apoyar la unidimensionalidad del autoconcepto. En conclusión, no hay apoyo empírico a la unidimensionalidad del autoconcepto, es más, como veremos a continuación, investigaciones posteriores han establecido claramente su multidimensionalidad, poniendo de manifiesto que el autoconcepto no puede ser adecuadamente comprendido si se ignora su naturaleza multidimensional (Byrne, 1984; Marsh y Shavelson, 1985; Shavelson y Marsh, 1986).

Se ha puesto de manifiesto la naturaleza multidimensional del autoconcepto, pero ¿qué relación guardan los diferentes factores que lo constituyen? Tanto el modelo multidimensional de factores independientes como el modelo multidimensional de factores correlacionados (ver figura 4.1.b y 4.1.c), representan el autoconcepto como un constructo multidimensional. La diferencia entre estos dos modelos radica en el grado en que las múltiples dimensiones del autoconcepto se encuentran correlacionadas. El modelo de factores independientes plantea que no existe ningún tipo de correlación entre los factores; aunque puede también formularse una versión menos restrictiva de este modelo que plantea la relativa ausencia de correlación entre los factores. Este modelo representa la antítesis del modelo unidimensional, ya que hipotetiza la no existencia de un factor general de autoconcepto, y aunque su versión más restrictiva no ha recibido apoyo por parte de ningún autor (Marsh y Hattie, 1996; Marsh, 1997), sin embargo, los trabajos de Soares y Soares (1977, 1983) y de Marsh y Shavelson (1985), pueden ser interpretados como un apoyo

empírico a la versión menos restrictiva del modelo. Soares y Soares (1977, 1983) encontraron los siguientes valores al correlacionar las nueve subescalas de su Inventario de Percepción Afectiva: correlaciones entre .14 y .66 (media r=.37) en una muestra de sujetos americanos, correlaciones entre .04 y .61 (media r=.28) en una muestra de sujetos italianos, y correlaciones entre .30 y .76 (media r=.57) en una muestra de sujetos españoles. Por su parte, Marsh y Shavelson (1985) ofrecen las correlaciones obtenidas entre las 13 escalas del SDQIII -"Self Description Questionnaire III"- (Marsh, 1992) en dos estudios realizados con estudiantes adolescentes, presentando éstas un valor medio de .09 y .08 respectivamente. Aunque los resultados de estos estudios indican una relativa falta de correlación entre los factores del autoconcepto, el modelo de factores correlacionados ha recibido mucho más apoyo empírico que el modelo de factores independientes (Marsh, 1997). La mayoría de los estudios han encontrado como mínimo correlaciones moderadas entre los diferentes factores del autoconcepto; e incluso en los estudios en los que las correlaciones encontradas han sido relativamente pequeñas, se ha encontrado evidencia de la existencia de una débil jerarquía en la estructura de las diferentes facetas del autoconcepto, lo que parece inconsistente con la lógica del modelo de factores independientes.

Otra forma de conceptualizar el autoconcepto es mediante un modelo multifacético (figura 4.1.d y 4.1.e). Marsh y Hattie (1996) explican la diferencia entre el modelo multidimensional y el modelo multidimensional multifacético utilizando como analogía el análisis de varianza de una vía y el manova. En el modelo multidimensional, que se correspondería con el anova de una vía, hay una única faceta (el contenido de los dominios del autoconcepto) que presenta múltiples

niveles (los diferentes dominios del autoconcepto, como por ejemplo, los dominios físico, social y académico). En el modelo multifacético taxonómico (ver figura 4.1.d), que se correspondería con el manova, hay como mínimo dos facetas, cada una de las cuales tiene dos o más niveles. La figura 4.1.e es una representación alternativa del modelo multifacético taxonómico, que se deriva del uso del análisis factorial confirmatorio (AFC) cuando el mismo o diferentes rasgos son medidos con métodos diferentes. Este modelo propone que todas las medidas que reflejan el mismo nivel de cada faceta pueden ser explicadas en términos de un único factor. En los estudios en los que se ha aplicado la metodología de matrices multirrasgo-multimétodo (MRMM), una de las facetas representa un dominio del autoconcepto, y la otra faceta representa diferentes métodos. Podemos citar como ejemplo el trabajo de Marsh, Byrne y Shavelson (1988) en el que los autores hicieron un análisis con matrices MRMM de los datos recogidos con tres instrumentos diferentes de medida del autoconcepto.

Soares y Soares (1977) desarrollaron un "modelo taxonómico" del autoconcepto tomando como base el modelo taxonómico de Guilford sobre la estructura de la inteligencia. Sin embargo, estos autores aunque proponen y defienden un "modelo taxonómico" del autoconcepto, dicha taxonomía parece quedarse en el nombre que recibe el modelo. Soares y Soares no definen su modelo con el suficiente detalle como para diferenciarlo del modelo de factores independientes expuesto anteriormente, y de hecho, aunque teóricamente proponen un "modelo taxonómico", empíricamente parece que estén evaluando un modelo de factores independientes.

Por otra parte, aunque son varios los instrumentos de medida del autoconcepto que de forma implícita o explícita asumen un modelo multifacético taxonómico (por ejemplo, la Escala de Autoconcepto de Tennessee -"Tennessee Self Concept Scale"- elaborada por Fitts (1965), o la Escala Multidimensional de Autoconcepto -"Multidimensional Self Concept Scale"- elaborada por Bracken (1992)), la forma de puntuar estos instrumentos parece inconsistente con el modelo que les subyace (Marsh y Richards, 1988; Marsh y Hattie, 1996). El modelo taxonómico puede permitir a los investigadores analizar de forma conjunta aspectos estructurales y aspectos de proceso del autoconcepto. Sin embargo, es necesario desarrollar futuras investigaciones que permitan determinar estructuras apropiadas para instrumentos que representen este modelo, así como la manera de puntuarlos para que ésta sea congruente con el modelo que les subyace (Marsh y Hattie, 1996).

Por último, el modelo multidimensional de factores jerárquicos (ver figura 4.1.f), incorpora en cierto modo cada uno de los modelos ya expuestos. Al igual que en el modelo unidimensional, se hipotetiza un componente global del autoconcepto que ocupa el punto más alto de la jerarquía. Por lo tanto, según Marsh y Hattie (1996), encontrar apoyo empírico del modelo unidimensional puede ser interpretado como evidencia a favor de un modelo jerárquico en el que la jerarquía es muy fuerte. En el extremo opuesto, encontrar apoyo empírico de la versión menos restrictiva del modelo multidimensional de factores independientes podría ser interpretado como evidencia a favor de un modelo jerárquico en el que la jerarquía es muy débil. Unicamente en el caso de que las correlaciones entre los factores del autoconcepto fueran de forma consistente iguales o próximas a cero, es decir, en caso de encontrar apoyo empírico a favor del modelo multidimensional de

factores independientes en su versión más restrictiva, la evidencia a favor del modelo jerárquico sería dudosa. Por otro lado, encontrar apoyo empírico del modelo multidimensional de factores correlacionados implica automáticamente evidencia a favor del modelo jerárquico. Finalmente, aunque la relación entre los modelos taxonómico y jerárquico no está muy claramente definida, parece que ambos modelos no son incompatibles (Marsh y Hattie, 1996).

A pesar de que ya William James (1890/1963) en sus primeras reflexiones sobre el autoconcepto resaltara su naturaleza multidimensional y jerárquica, la investigación empírica anterior a 1980 se centró únicamente en el estudio de un autoconcepto global. Además, estudios de revisión de las investigaciones llevadas a cabo antes de la década de los 80 (Burns, 1979; Shavelson, Hubner y Stanton, 1976; Wells y Marwell,1976; Wylie, 1974, 1979) ponen de manifiesto la ausencia de bases teóricas en la mayoría de los estudios, la pobre calidad de los instrumentos de medida utilizados para evaluar el autoconcepto, la presencia de deficiencias metodológicas, y una carencia general de consistencia en los resultados encontrados. Sin embargo, los estudios realizados a partir de los años 80, han representado un importante avance en la teorización, medida e investigación del autoconcepto. El trabajo de Shavelson, Hubner y Stanton (1976) representó un importante punto de partida para el desarrollo de dichos avances. Estos autores señalan el renovado y creciente interés de los investigadores hacia el estudio del autoconcepto, pero critican también la existencia de importantes deficiencias en la investigación de este constructo. A partir de esta crítica, su aportación consiste en ofrecer un modelo del autoconcepto en el que se resalta su naturaleza multidimensional y jerárquica. Dicho modelo (que se correspondería con el representado en

la figura 4.1.f) resultó ser la base para el desarrollo posterior de una nueva generación de instrumentos multidimensionales de medida del autoconcepto que han representado un importante avance en el desarrollo de la investigación de este constructo. Tal como veremos más adelante, el cuestionario PSDQ -"Physical Self Description Questionnaire"- (Marsh, Richards, Johnson, Roche y Tremayne, 1994), objeto de estudio en este trabajo, tiene también sus raíces en este modelo, por lo que vamos a pasar a desarrollarlo en el siguiente apartado.

1.3. El modelo multidimensional jerárquico de Shavelson y cols.

Shavelson, Hubner y Stanton (1976) hicieron una revisión de las diferentes definiciones del autoconcepto que aparecían en la literatura, y tras integrar los aspectos comunes y diferenciales existentes entre ellas, desarrollaron su propia definición. Estos autores definen el autoconcepto como las percepciones que tiene el individuo sobre sí mismo, que se forman a través de la experiencia y de la interpretación de su entorno. Estas percepciones se ven fuertemente influidas por las evaluaciones de los otros significativos.

Shavelson y cols. identificaron siete aspectos fundamentales que definen el autoconcepto tal y como ellos lo conceptualizan:

1° Es una estructura organizada: en base a su autoconcepto, el individuo categoriza la gran cantidad de información que tiene sobre sí mismo y establece relaciones entre dichas categorías.

2° Es multidimensional: presenta dimensiones claramente diferenciadas.

3° Es jerárquico: las percepciones de la conducta personal en situaciones específicas se encuentran en la base de dicha jerarquía, las inferencias sobre uno mismo en dominios más amplios (por ejemplo el dominio social, físico o académico) ocupan la parte media, y finalmente, un autoconcepto general y global ocupa la parte superior de dicha jerarquía.

4° El autoconcepto global (que ocupa la parte superior de la jerarquía) es estable, pero conforme se desciende en dicha jerarquía, el autoconcepto se vuelve más específico y dependiente de las situaciones, y por lo tanto menos estable.

5° El autoconcepto aumenta su multidimensionalidad con la edad: los bebes no diferencian entre ellos mismos y su entorno; los niños presentan un autoconcepto global, no diferenciado, y específico de cada situación; al aumentar la edad del niño y sobre todo con la adquisición del lenguaje que le permite utilizar etiquetas verbales, desarrolla de forma progresiva un autoconcepto más diferenciado, integrado por diferentes dimensiones y que presenta una estructura jerárquica.

6° El autoconcepto, como percepción que el individuo tiene sobre sí mismo, presenta tanto aspectos descriptivos como aspectos evaluativos: el individuo puede describirse a sí mismo (por ejemplo, "Soy alto"), y también puede evaluarse (por ejemplo, "Soy bueno jugando al baloncesto"). Esta evaluación puede ser hecha en base a un ideal a alcanzar, la comparación con los compañeros, o las expectativas de los

otros significativos. Al igual que ya señaló William James, estos autores indican que el individuo puede ponderar de forma diferencial la importancia dada a las diferentes dimensiones.

7° El autoconcepto representa un constructo con entidad propia: puede ser claramente diferenciado de otros constructos con los cuales está teóricamente relacionado.

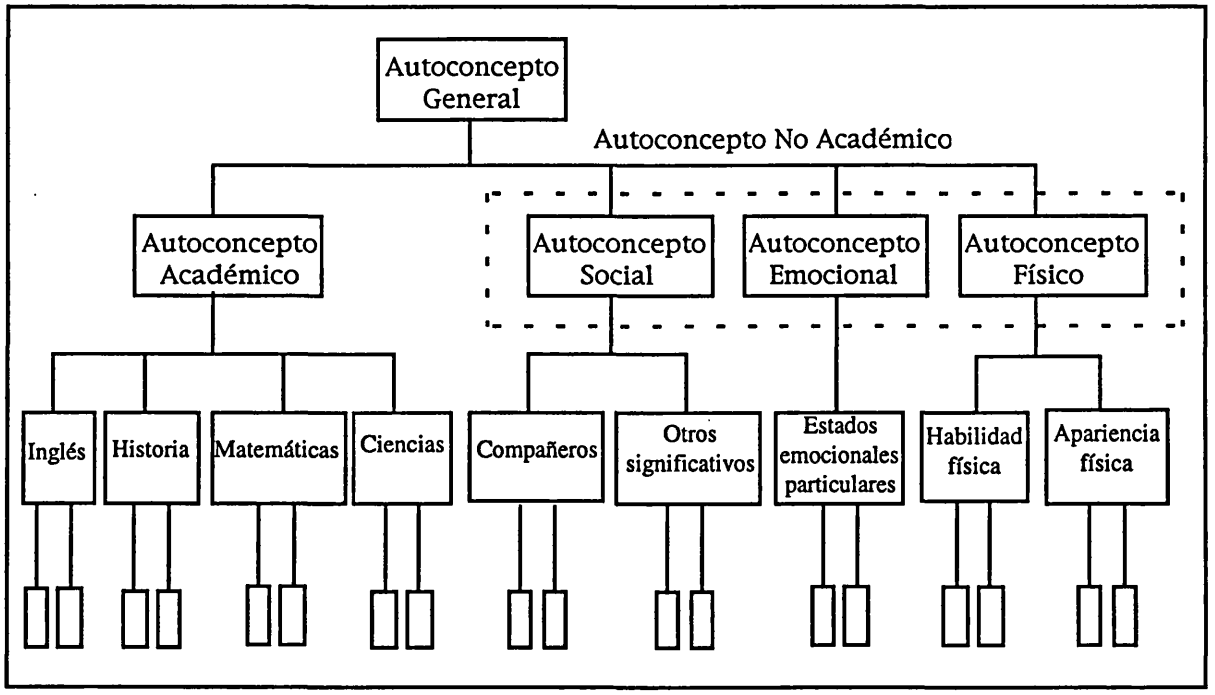


Fig. 4.2. Representación del modelo multidimensional jerárquico del autoconcepto propuesto por Shavelson y cols. (1976)

Shavelson y cols. (1976) ofrecen una posible representación de su modelo multidimensional jerárquico (ver figura 4.2.). Este modelo presenta un autoconcepto general en la parte superior de la jerarquía, que aparece dividido en el nivel inferior en autoconcepto académico y autoconcepto no académico. El autoconcepto no académico se divide a su vez en social, emocional y físico. Es decir, el autoconcepto global es el resultado de las percepciones que el sujeto tiene de sí mismo en los dominios académico y no académico (social, emocional y físico). Cada uno

de los dominios en los que se divide el autoconcepto global aparecen a su vez divididos en diferentes subdominios. El autoconcepto académico se divide en autoconceptos para particulares materias académicas; el autoconcepto social se divide en relaciones con los compañeros (pares o iguales), y relaciones con los otros significativos; y el autoconcepto físico se divide en habilidad física y apariencia física). De este modo, por ejemplo, el autoconcepto académico es el resultado de las percepciones de las habilidades específicas del sujeto en materias académicas concretas (inglés, historia, matemáticas, ciencias); y de forma paralela, el autoconcepto físico es el resultado de la combinación de las percepciones del sujeto sobre sí mismo en cuanto a habilidad física y apariencia física. Por debajo de estos subdominios aparecen otros niveles que representan percepciones mucho más específicas y dependientes de situaciones concretas.

El modelo de Shavelson y cols. (1976) ha sido objeto de revisiones posteriores que han supuesto su ampliación o la modificación de algunos de los aspectos de la definición del autoconcepto tal como aparece en el modelo original. Sin embargo, la aportación fundamental de estos autores, es decir, su definición del autoconcepto desde un modelo multifacético y jerárquico, ha recibido amplio apoyo empírico (Shavelson y Bolus, 1982; Byrne, 1984; Fleming y Courtney, 1984; Fleming y Watts, 1980; Marsh, 1987b; Marsh y Shavelson, 1985; Shavelson y Marsh, 1986).

Alguno de los aspectos que han sido puesto en entredicho es el postulado en el punto 4° sobre la estabilidad del autoconcepto. Shavelson y Bolus (1982) examinaron este supuesto que señala la creciente inestabilidad del autoconcepto conforme se baja en la jerarquía

de niveles. En este estudio analizaron datos sobre el autoconcepto de escolares recogidos en dos ocasiones diferentes separadas por un intervalo de cinco meses, utilizando una cuestionario que abarcaba diferentes niveles de este constructo (autoconcepto general, autoconcepto académico, inglés, matemáticas y ciencias). Los resultados obtenidos indicaron una gran estabilidad de los subdominios específicos del autoconcepto (inglés, matemáticas, y ciencias), incluso mayor que la estabilidad presentada por el autoconcepto académico. Otros estudios posteriores (Marsh, 1990a, Byrne, 1984) también han puesto en entredicho que el autoconcepto global sea el componente más estable de la estructura jerárquica.

Posteriormente, Marsh y Shavelson (Marsh y Shavelson,1985; Shavelson y Marsh, 1986), llevaron a cabo otra revisión del modelo original de Shavelson y cols. Estos autores pusieron a prueba el ajuste de tres modelos diferentes: un primer modelo con un único factor de primer orden (autoconcepto general); un segundo modelo que además incluye dos factores independientes de segundo orden (autoconcepto académico y autoconcepto no académico); y un tercer modelo que incluye un factor de primer orden (autoconcepto general) y tres factores independientes de segundo orden (autoconcepto académico verbal, autoconcepto académico matemático, y autoconcepto no académico). El segundo modelo presentaba mejor ajuste que el primero, sin embargo, fue el tercer modelo el que presentaba el mejor ajuste. En base a estos resultados, Marsh y Shavelson, resaltando el apoyo que representan sus estudios a la conceptualización del autoconcepto como un constructo multidimensional y jerárquico, proponen sin embargo una pequeña modificación estructural del modelo original de Shavelson y cols. La aportación de estos autores puede verse en la figura 4.3.

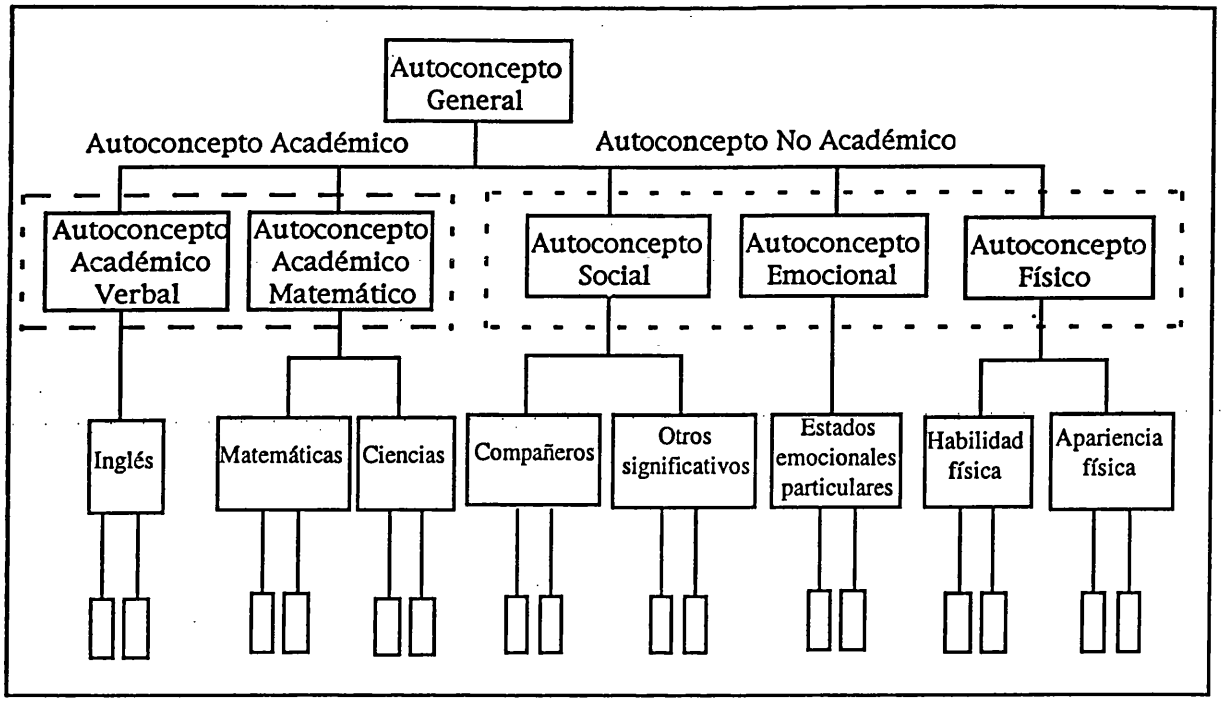


Fig. 4.3. Modelo Marsh/Shavelson (1985): revisión del modelo multidimensional jerárquico de Shavelson y cols. (1976).

El modelo de Shavelson y cols. (1976) ha servido de base para el desarrollo de otros modelos multidimensionales del autoconcepto (Song y Hattie, 1984; Hattie, 1992; Harter, 1982; Pallas, Entwisle, Alexander, y Weinstein, 1990). Sin embargo, como ya hemos comentado anteriormente, su principal aportación ha sido fomentar el desarrollo de instrumentos de medida que recogen la naturaleza multidimensional y jerárquica del autoconcepto. De la medición del autoconcepto hablaremos en el siguiente apartado.

2. LA MEDICION DEL AUTOCONCEPTO

En el apartado anterior hemos expuesto algunas pinceladas históricas de la evolución en la conceptualización del autoconcepto. Hemos visto cómo se ha pasado de un modelo unidimensional que postulaba un único factor general, a un modelo multidimensional en el

que los diferentes componentes del autoconcepto se disponen de forma jerárquica. Esta evolución teórica en la definición del autoconcepto, se ha visto reflejada de forma paralela en el desarrollo de instrumentos para su medición.

2.1. Evolución en los instrumentos de medida del autoconcepto

Los primeros autores que desarrollaron instrumentos para medir el autoconcepto lo hicieron desde el modelo dominante en aquel momento, el modelo unidimensional. Por lo tanto, estas primeras escalas estaban compuestas por un conjunto de items en los que se pedía al sujeto que se valora a sí mismo en una amplia variedad de cualidades y habilidades personales, y en un amplio rango de situaciones. Todos los items eran valorados de la misma manera, y la puntuación total, obtenida como la suma de las puntuaciones individuales en cada uno de los items, representaba el nivel de autoconcepto global del individuo.

Entre estas primeras escalas que ofrecen una medida unidimensional del autoconcepto podemos citar: la Escala de Autoestima de Rosenberg -"Rosenberg Self Esteem Scale"- (Rosenberg, 1965, 1979), los Inventarios de Autoestima de Coopersmith -"Coopersmith Self Esteem Inventories"- (Coopersmith, 1967), la Escala de Autoconcepto de Tennessee -"Tennessee Self Concept Scale"- (Fitts, 1965), y la Escala de Autoconcepto de Piers-Harris -"Piers-Harris Children's Self Concept Scale"- (Piers, 1969, 1984). Sin embargo, la utilidad de estas escalas para el estudio del autoconcepto queda en entredicho cuando desde un punto de vista teórico se comienza de defender la necesidad de considerar la naturaleza multidimensional del autoconcepto. Como resultado de esta evolución teórica, la mayoría de los instrumentos que se desarrollaron a

partir de la segunda mitad de los años 70 para medir el autoconcepto, evaluaban diferentes dominios de este constructo además de un componente global. Incluso algunos de los primeros instrumentos desarrollados en la etapa anterior y que por tanto ofrecían medidas unidimensionales del autoconcepto, fueron objeto de revisiones y modificaciones para incorporar los avances teóricos en la estructura del autoconcepto y ofrecer de este modo una medida multidimensional del autoconcepto. Tal es el caso de las revisiones de los Inventarios de Autoestima de Coopersmith (Coopersmith, 1981), y de la Escala de Autoconcepto de Tennessee (Roid y Fitts, 1988).

Son muchas las escalas que se desarrollan desde la formulación multidimensional del autoconcepto. Todas ellas recogen diferentes dimensiones de este constructo. Sin embargo, tal como hemos visto al exponer los modelos que contemplan la naturaleza multidimensional del autoconcepto (modelo multidimensional de factores independientes, modelo multidimensional de factores correlacionados, y modelo multidimensional de factores jerárquicos), las dimensiones particulares incluidas en cada instrumento, la estructura exacta de estas dimensiones, y la relación entre ellas y de éstas con el autoconcepto global, varía ampliamente en los diferentes instrumentos.

El Cuestionario Revisado de Autoimagen de Offer -"Offer Self Image Questionnaire Revised"- (Offer, Ostrov, Howard, y Dolan, 1992), el Perfil de Autopercepción para Niños -"Self Perception Profile for Children"- (Harter, 1982, 1985), el Indice de Autoestima -"Self Esteem Index"- (Brown y Alexander, 1991), son algunos ejemplos de instrumentos que parten de la conceptualización del autoconcepto como un constructo multidimensional.

Otro de los modelos comentados para explicar la estructura del autoconcepto es el modelo multifacético taxonómico. Ya hemos indicado anteriormente alguna de las escalas a las que subyace dicho modelo. Brevemente, vamos a describir la estructura de una de ellas, para ejemplificar cómo se formula un modelo taxonómico y qué aspectos adicionales permite evaluar respecto a los modelos multidimensionales. Por ejemplo, la Escala de Autoconcepto de Tennessee (Fitts, 1965), aunque de forma explícita parte de la base teórica de un modelo unidimensional de autoconcepto, implícitamente asume un modelo multifacético taxonómico con tres facetas diferenciadas (Marsh y Hattie, 1996). El diseño que plantea Fitts se define en los siguientes términos: 5 (marco de referencia externo) X 3 (marco de referencia interno) X 2 (items enunciados positiva o negativamente). Es decir, presenta 3 facetas con diferentes niveles en cada una de ellas. La primera faceta, denominada faceta externa o marco de referencia externo, presenta 5 niveles (físico, moral, personal, familiar, y social), que se corresponderían con los dominios del autoconcepto evaluados desde un modelo multidimensional simple. La aportación del modelo taxonómico es que cada uno de estos dominios puede ser evaluado en relación con los niveles de otras facetas relevantes. En el esquema de Fitts, la segunda faceta, denominada marco de referencia interno, presenta 3 niveles: Identidad (por ejemplo, "qué soy"), Satisfacción (por ejemplo, "cómo me siento sobre mí mismo"), y Conducta (por ejemplo, "qué hago o cómo me comporto"). La Identidad representa un autoconcepto interno y privado; la Conducta representa el self observable por los otros; y la Satisfacción refleja la discrepancia entre ideal y realidad. Por último, la tercera faceta representa la forma en la que están enunciados los items (positiva o negativamente), y carece de importancia sustantiva, ya que su único propósito es controlar ciertos sesgos de respuesta a los items.

Hemos visto el paralelismo que se establece entre la evolución en la conceptualización teórica y el desarrollo de instrumentos para medir el autoconcepto: la asunción de diferentes modelos teóricos implica desarrollar instrumentos de medida que permitan poner a prueba dichos modelos. Como ya ha sido comentado anteriormente, uno de los modelos multidimensionales más ampliamente investigado, y que ha originado el desarrollo de numerosos instrumentos de medida del autoconcepto, es el modelo de Shavelson y cols. (1976). Algunas de las escalas que toman como punto de partida este modelo multidimensional jerárquico del autoconcepto son: la Escala Multidimensional de Autoconcepto -"Multidimensional Self Concept Scale"- (Bracken, 1992), la Escala de Percepción de Habilidad para Estudiantes -"Perception of Ability Scale for Students"- (Boersma y Chapman, 1992), y las tres versiones del Cuestionario de Autopercepción -"Self Description Questionnaire"- SDQI, SDQII y SDQIII (Marsh,1988, 1990b,1992).

En el siguiente apartado vamos a centrar nuestra atención en la descripción de estas tres últimas escalas, ya que suponen el punto de partida del PSDQ, el cuestionario que va a ser analizado en el presente trabajo. Para una descripción más detallada de las otras escalas citadas en este apartado, remitimos al lector interesado al capítulo de Keith y Bracken (1996). Estos autores presentan una revisión detallada de los principales instrumentos de medida del autoconcepto, ofreciendo información sobre su desarrollo y estandarización, el procedimiento de administración y de interpretación de las puntuaciones, y sus principales características psicométricas.

2.2. Desarrollo de las diferentes versiones del SDQ

El estudio de cualquier constructo psicológico requiere partir de conceptualizaciones teóricas que deben ser medidas adecuadamente. Estas operacionalizaciones de los constructos deben ser evaluadas empíricamente, es decir, es necesario investigar su validez de constructo. Las investigaciones que pretenden validar las operacionalizaciones de los constructos pueden ser clasificadas como estudios de validación interna y estudios de validación externa. En los primeros, el interés radica en explorar la estructura interna del constructo; en los segundos, se intenta establecer el tipo de relación que existe entre el constructo de interés y otros constructos. Parece evidente que la validación de un constructo requiere comenzar con el estudio de su estructura interna. Sin embargo, la investigación sobre el autoconcepto llevada a cabo antes de la década de los 80, se centró predominantemente en la validación externa de este constructo, es decir, la mayoría de los investigadores estaban interesados en estudiar las relaciones entre el autoconcepto y otros constructos. El desarrollo posterior de modelos que plantean la multidimensionalidad del autoconcepto supuso un empuje en la proliferación de estudios sobre la validez interna de este constructo. Se desarrollan instrumentos de medida que contemplan las diferentes dimensiones del autoconcepto postuladas desde el modelo teórico, y se analiza la validez del modelo mediante técnicas empíricas como el análisis factorial o el análisis de matrices multirrasgo-multimétodo.

A partir del modelo de Shavelson y cols., y para poner a prueba su validez interna, se desarrolló el cuestionario de autoconcepto SDQ -"Self Description Questionnaire"- (Marsh, 1986). Este cuestionario evalúa 3 áreas del autoconcepto académico, 4 áreas del autoconcepto no

académico, y el autoconcepto global. A continuación presentamos una breve descripción de las 8 subescalas que componen el cuestionario, indicando qué dominio del autoconcepto representan según el modelo de Shavelson y cols.:

Autoconcepto Académico:

1) Lectura: percepción de los estudiantes sobre su habilidad y satisfacción/interés en la lectura;

2) Matemáticas: evaluación de los estudiantes sobre su habilidad y satisfacción/interés en matemáticas;

3) Escolar: percepción de los estudiantes sobre su habilidad y satisfacción/interés en todas las asignaturas escolares en general;

Autoconcepto Social (no académico):

4) Relación con los compañeros: percepción de los estudiantes sobre su facilidad en hacer amigos, su popularidad, y si los otros lo quieren como amigo;

5) Relación con los padres: percepción de los estudiantes sobre el nivel de la relación con sus padres (facilidad para comunicarse con ellos, comprensión por parte de los padres, percepción de afecto y satisfacción hacia él);

Autoconcepto Físico (no académico):

6) Habilidad física: percepción de los estudiantes sobre su habilidad y satisfacción/interés en actividades físicas, deportes y juegos;

7) Apariencia física: percepción de los estudiantes sobre su propio atractivo, su apariencia física en comparación con los otros, y su percepción sobre lo que los otros piensan de su apariencia;

Autoconcepto General:

8) Autoconcepto general: percepción global de los estudiantes sobre sí mismos, y su satisfacción y orgullo respecto a cómo se perciben. Esta subescala aunque no figuraba en la formulación inicial del SDQ, fue posteriormente añadida, y se basa en la Escala de Autoestima de Rosenberg (1965).

Posteriormente se desarrollaron diferentes versiones de este instrumento para evaluar el autoconcepto en diferentes rangos de edad que representan la adolescencia temprana, media y tardía. El SDQI (Marsh, 1988) presenta la estructura descrita en la formulación original del instrumento (Marsh, 1986), y se define como adecuado para medir múltiples dimensiones del autoconcepto en sujetos de edades comprendidas entre 8 y 12 años (adolescencia temprana).

El SDQII (Marsh, 1990b) se formula para medir el autoconcepto en sujetos de edades comprendidas entre 12 y 16 años (adolescencia media). Su estructura presenta alguna modificación con respecto a la versión anterior: la subescala "Relación con los compañeros" se divide en dos subescalas diferenciadas denominadas "Relación con los compañeros del mismo sexo" y "Relación con los compañeros del sexo contrario"; además, el SDQII incluye dos nuevas subescalas "Estabilidad emocional" (percepción de los estudiantes sobre ellos mismos como personas tranquilas, nerviosas, que se preocupan con facilidad, etc.) y "Honradez-Sinceridad" (percepción de los estudiantes sobre su honradez, sinceridad, y como personas dignas de confianza). Otra modificación hace referencia a la nomenclatura utilizada para definir una de las subescalas del autoconcepto académico: la subescala "Lectura" recibe el nombre de "Verbal" en esta nueva versión. Suponemos que esta modificación

pretende recoger con mayor claridad la reformulación, comentada en páginas anteriores, que el modelo de Marsh/Shavelson plantea respecto al modelo original de Shavelson y cols.

Por último, el SDQIII (Marsh, 1992) es adecuado para medir el autoconcepto en sujetos a partir de los 17 años (adolescencia tardía y jóvenes adultos). Ha sido utilizado principalmente con estudiantes universitarios (Marsh y Byrne, 1993a, 1993b), e incluye dos nuevas escalas con respecto a la versión anterior, "Valores espirituales/Religión" (percepción de los estudiantes sobre ellos mismos como personas religiosas, y la importancia que las creencias religiosas tienen en el desarrollo de su vida) y "Resolución de problemas" (percepción de los adolescentes sobre su habilidad para resolver problemas y pensar de forma creativa e imaginativa). En la figura 4.4. aparece un esquema de las subescalas que componen estas tres escalas, así como una indicación de los rangos de edad y cursos para los que han sido diseñadas.

Además de contar con un respaldo teórico importante, las diferentes versiones del SDQ, en base a los estudios sobre sus propiedades psicométricas y a la evidencia de su validez interna (Boyle, 1994; Byrne, 1984; Hattie, 1992; Wylie, 1989), han demostrado ser uno de los instrumentos multidimensionales más adecuados para la medida del autoconcepto (Marsh, 1997). Por otra parte, el desarrollo de estos cuestionarios representa un claro ejemplo de interacción entre fundamentación teórica e investigación empírica. El modelo de Shavelson y cols. fue tomado como punto de partida para la construcción de estos instrumentos, y los resultados empíricos obtenidos a partir de ellos, han sido utilizados para apoyar y en algunos casos revisar o modificar el modelo teórico del que partían, lo que ha resultado de forma paralela en

modificaciones de los instrumentos utilizados para poner a prueba las formulaciones teóricas.

SDQI	SDQII	SDQIII
Autoconcepto General: 1) General Autoconcepto Académico: 2) Lectura 3)Matemáticas 4) Escolar Autoconcep.No Académico: 5) Relación con compañeros 6) Relación con padres 7) Habilidad física 8) Apariencia física	Autoconcepto General: 1) General Autoconcepto Académico: 2) Verbal 3)Matemáticas 4) Escolar Autoconcep.No Académico: 5) Relación con compañ. del mismo sexo 6) Relación con compañ. del sexo contrario 7) Relación con padres 8) Habilidad física 9) Apariencia física 10) Estabilidad emocional 11) Honradez-Sinceridad	Autoconcepto General: 1) General Autoconcepto Académico: 2) Verbal 3)Matemáticas 4) Escolar 5) Resolución problemas Autoconcep.No Académico: 6) Relación con compañ. del mismo sexo 7) Relación con compañ. del sexo contrario 8) Relación con padres 9) Habilidad física 10) Apariencia física 11) Estabilidad emocional 12) Honradez-Sinceridad 13) Religión
8- 12 años	12-16 años	A partir 17 años
4° a 6° Primaria	1° a 4° E.S.O.	1°-2° Bachillerato y Universidad

Fig. 4.4. SDQI, SDQII y SDQIII: subescalas que las componen, rangos de edad y cursos para los que han sido diseñadas.

Los estudios llevados a cabo con las diferentes versiones del SDQ para analizar la estructura del autoconcepto han utilizado muestras de estudiantes de diferentes edades. Por tanto, no es de extrañar que uno de los dominios que ha acaparado mayor atención haya sido el autoconcepto académico. Esto, unido al intento de proporcionar mayor evidencia a favor del modelo revisado de Marsh/Shavelson (1985), llevó a la elaboración del Cuestionario de Autodescripción Académica ASDQ - "Academic Self Description Questionnaire"- (Marsh, 1990c). Este instrumento mide 15 dimensiones del autoconcepto académico, que representan otras tantas asignaturas escolares diferentes: lengua inglesa, literatura inglesa, lengua extranjera, historia, geografía, comercio, informática, ciencias, matemáticas, educación física, salud, música, arte, arte industrial, y religión. Un estudio de la estructura factorial de este instrumento (Marsh, 1990c) puso de manifiesto la existencia de 15 factores de primer orden correspondientes a las subescalas del cuestionario. Sin embargo, la formulación de dos factores de orden superior (autoconcepto académico verbal y autoconcepto académico matemático) no era suficiente para explicar las relaciones existentes entre los factores de primer orden (éstos únicamente eran capaces de explicar las relaciones entre ocho asignaturas troncales), por lo que fue necesario formular factores adicionales de orden superior para explicar las relaciones entre otras asignaturas no troncales (por ejemplo, educación física, arte y música).

La medida del autoconcepto pasó de instrumentos unidimensionales que medían un constructo global, a instrumentos multidimensionales que ponen de manifiesto la necesidad de considerar las diferentes dimensiones del autoconcepto. Sin embargo, el desarrollo del ASDQ y los estudios realizados con él, llevaron a Marsh a reflexionar

y a argumentar la necesidad de ir un poco más lejos, es decir, de elaborar instrumentos de medida específicos para cada una de las dimensiones del autoconcepto (académico, social, físico, etc.); solamente de este modo es posible profundizar en el estudio de la estructura del autoconcepto. El ASDQ ya había sido un primer paso en este sentido en el dominio del autoconcepto académico. El PSDQ -"Physical Self Description Questionnaire"- (Marsh, Richards, Johnson, Roche y Tremayne, 1994) pretende ser una aportación similar para profundizar en el estudio de la estructura del autoconcepto físico. En el siguiente apartado vamos a tratar esta dimensión del autoconcepto.

3. EL AUTOCONCEPTO FISICO

El autoconcepto físico ha sido definido como las percepciones que tienen los sujetos sobre sus habilidades físicas y su apariencia física (Stein, 1996). Desde modelos multidimensionales del autoconcepto, como el formulado por Shavelson y cols. (1976), el autoconcepto físico representa una de las dimensiones que contribuyen a la formación de un autoconcepto global. Además, estudios realizados sobre la estructura interna de este constructo, ofrecen también evidencia empírica en favor de su multidimensionalidad. Franzoi y Shields (1984) informaron de tres dimensiones del autoconcepto físico: habilidad física, apariencia física, y conductas de control de peso. Otros autores (Marsh 1986; Marsh y Jackson, 1986; Marsh y Peart, 1988) han diferenciado únicamente dos dimensiones: habilidad física y apariencia física.

Diferentes estudios ponen de manifiesto la contribución del autoconcepto físico en el desarrollo del autoconcepto global. Dicha

contribución es evidente desde la más tierna infancia, ya que según parece, el autoconcepto físico representa uno de los componentes más importante del autoconcepto global en los niños (Fisher y Cleveland, 1968; Klesges, Haddock, Stein, Klesges, Eck y Hanson, 1992; Keller, Ford, y Meacham, 1978). Por ejemplo, el estudio de Fisher y Cleveland (1968) pone de manifiesto que la insatisfacción de los niños con sus propios cuerpos presenta importantes efectos negativos sobre su autoconcepto global.

Por otra parte, estudios realizados con adultos ofrecen también evidencia a favor de la citada contribución: de forma consistente, se han encontrado correlaciones positivas entre apariencia física y autoconcepto global en adultos (Adams, 1977; Lerner y Karabenick, 1974; Lerner, Karabenick, y Stuart, 1973; Lerner, Orlos, y Knapp,1976; Mathes y Kahn, 1975; Simmons, y Rosenberg, 1975). Hatfield y Sprecher (1986) argumentan que estas diferencias en el desarrollo del autoconcepto son debidas al trato diferencial que reciben de la sociedad los sujetos atractivos con respecto a los sujetos menos atractivos. Según estos autores, los adultos atractivos y que poseen mayores habilidades físicas suelen tener mayor popularidad, recibir un trato preferencial, y ser más frecuentemente alabados por sus cualidades físicas que los adultos menos atractivos.

Pero quizá sea en la adolescencia donde el autoconcepto físico cobra más relevancia. Esta es una etapa de importantes cambios físicos en la que se desarrollan los caracteres sexuales secundarios. Debido a estos cambios, la apariencia física se convierte en una de las preocupaciones centrales en la vida de los adolescentes (Clifford, 1971). Este autor presenta evidencia de que aunque los adolescentes por norma

general suelen percibir su apariencia física de forma positiva, entre sus principales preocupaciones se encuentran el físico, el peso y la altura. Otro estudio que pone de manifiesto la relevancia que adquieren los aspectos físicos para el adolescente, es el realizado por Newcomb y Bukowski (1983). En este trabajo se analizan los aspectos que valoran niños y adolescentes de diferentes edades a la hora de elegir a sus amigos. Los resultados indican que conforme los niños se acercan a la adolescencia, valoran más la apariencia física y las habilidades deportivas frente a las habilidades académicas a la hora de elegir a sus amigos.

Dada la importancia que la apariencia y las habilidades físicas percibidas adquieren para el adolescente, se puede hipotetizar que la participación en actividades que permitan potenciar las habilidades físicas y mejorar la apariencia física, facilitará el desarrollo de un autoconcepto físico más positivo. Jackson y Marsh (1986) encontraron correlaciones positivas entre la participación en actividades deportivas y la habilidad física percibida en una muestra de mujeres adolescentes y jóvenes. Por su parte, Marsh y Peart (1988) encontraron que la participación en actividades aeróbicas tenía un importante efecto sobre la habilidad física percibida; estos autores informan de correlaciones positivas altas entre la forma física y la habilidad física percibida, mientras que las correlaciones entre la forma física y la apariencia física fueron también positivas pero moderadas. Sin embargo, también aparece en la literatura algún estudio que no apoya la hipótesis planteada. Bakker (1988) estudió el autoconcepto físico en mujeres adolescentes que asistían a cursos de baile organizados en la escuela. En base a los resultados anteriores, parece intuitivo hipotetizar que asistir a una actividad como las clases de baile podría afectar de forma positiva

el autoconcepto físico de estas adolescentes. Sin embargo Bakker encontró justo lo contrario: las adolescentes que asistían a los cursos de baile presentaban un autoconcepto físico más bajo que aquellas que no asistían, tanto en habilidades físicas como en apariencia física. Bakker intenta explicar estos resultados en base al ambiente competitivo de la escuela y a las evaluaciones críticas que se producen en este ambiente respecto a la apariencia y las habilidades físicas. Este estudio nos parece interesante y nos hace reflexionar sobre lo contraproducente que puede llegar a ser el sugerir a un/a adolescente preocupado/a por su peso o por su apariencia física apuntarse a un gimnasio, donde además de poder mejorar su condición física, tendrá ocasión de disfrutar de un ambiente crítico y obsesivo por la perfección corporal, y donde probablemente, lejos de resolver su problema, tendrá ocasión de agudizarlo.

Dejando aparte esta última reflexión, y teniendo en cuenta que las características del entorno sean las idóneas, otro de los grupos poblacionales que puede verse beneficiado de participar en actividades que potencien las habilidades físicas son los sujetos que presentan algún tipo de discapacidad física. Teniendo en cuenta la relación existente entre el autoconcepto físico y el autoconcepto global ya desde la infancia, y que la habilidad física es uno de los componentes del autoconcepto físico, puede hipotetizarse que si un niño presenta un funcionamiento físico limitado en algún aspecto, esta limitación puede afectar negativamente su autoconcepto físico, y por lo tanto disminuir de forma proporcional su autoconcepto global (Harvey y Greenway, 1984; Stein, 1996). De hecho, la práctica clínica ha documentado la alta incidencia de depresión y de baja autoimagen en niños con discapacidades físicas (Schecter, 1961).

Respecto a las diferencias de género, la literatura sugiere que las chicas suelen estar más preocupadas por los aspectos físicos que los chicos, sobre todo respecto a la apariencia física y al control de peso (Collins, 1991; Garner y cols., 1980). Sin embargo, en sujetos superdotados parece que esta relación se invierte y son los chicos los que muestran mayor preocupación por la apariencia física (Cornell y cols. 1990).

Otro de los aspectos que aparece relacionado con el autoconcepto físico son los trastornos de la conducta alimenticia. Diferentes investigaciones ponen de manifiesto que la percepción que los niños tienen de su apariencia física, así como su autoconcepto físico global, están estrechamente relacionados con el desarrollo de trastornos de la conducta alimenticia (Abraham y Beaumont, 1982; Loro y Orleans, 1981; Pyle, Mitchell, y Lokert, 1981). Es evidente la importancia que se da hoy en día a la apariencia física y al cuidado del cuerpo, y la presión social que reciben al respecto los niños y los adolescentes en su vida cotidiana (publicidad, televisión, etc.). Collins (1991) realizó un estudio con niños de 6 a 9 años de edad. Se pidió a los niños que eligieran, entre diferentes dibujos que representaban la figura humana, aquél que mejor les representaba en la actualidad, y aquél que representaba su ideal sobre cómo les gustaría ser. Los resultados mostraron que el 42% de las niñas y el 30% de los niños indicaron figuras ideales más delgadas que su representación actual. Por su parte, Stein y cols. (1995) encontraron que el 41% de los niños con edades comprendidas entre 6 y 11 años que configuraban la muestra de su estudio, deseaban cambiar las dimensiones de su cuerpo. Además, estos autores encontraron que el 25% de los niños de menor edad informaron sobre la realización de conductas para modificar su peso. Maloney y cols. (1988) encontraron

que cerca del 7% de los sujetos de una muestra de niños de 8 a 12 años, puntuaron dentro del rango de anorexia nerviosa en el Test de Actitudes Alimenticias para Niños (ChEAT) -"Children Eating Attitudes Tests"-. Todos estos resultados apuntan hacia la necesidad de profundizar en el estudio de la relación existente entre el autoconcepto físico y el desarrollo de trastornos de la conducta alimenticia (Stein, 1996).

4. LA MEDICION DEL AUTOCONCEPTO FISICO

Una vez puesta de manifiesto la importancia del estudio del autoconcepto físico, vamos a pasar a tratar el tema de la medición de este constructo. En primer lugar expondremos una visión general de la evolución en la medida del autoconcepto físico, que corre de forma paralela y enlaza con lo comentado en el apartado anterior respecto a la medición del autoconcepto general. Presentaremos también algunos de los instrumentos de medida del autoconcepto que aparecen en la literatura como más adecuados para la medida de este constructo. Y finalmente pasaremos a describir el PSDQ, su proceso de construcción, y los resultados de los estudios llevados a cabo para analizar sus propiedades psicométricas, su estructura interna y su validez de constructo.

4.1. Evolución en la medida del autoconcepto físico

Como ya hemos visto, los primeros cuestionarios que se desarrollaron para medir el autoconcepto lo hicieron desde una conceptualización unidimensional de este constructo. De este modo, aunque muchos de ellos contenían items que hacían referencia a habilidades físicas y a elementos de la apariencia física, no aparecía una

dimensión diferenciada que fuera interpretada como medida del autoconcepto físico. Una excepción es la Escala de Autoconcepto de Tennessee (Fitts, 1965), que aunque de forma explícita asume una conceptualización unidimensional del autoconcepto, de forma implícita representa diferentes dominios de este constructo, entre ellos el físico. La subescala de autoconcepto físico de este cuestionario ha sido ampliamente utilizada en las primeras investigaciones llevadas a cabo sobre la relación entre la práctica deportiva y el autoconcepto físico. Sin embargo, esta subescala recoge autopercepciones en áreas tan diversas (salud, pulcritud de la apariencia, atractivo físico, forma física), que la combinación de todas en ellas en una única escala hace poner en duda la validez de dicha escala (Marsh y Richards, 1988).

Posteriormente, la conceptualización teórica multidimensional del autoconcepto dio como resultado el desarrollo de instrumentos que pretendían medir las diferentes dimensiones de este constructo. En la figura 4.5. aparecen algunos de los cuestionarios multidimensionales del autoconcepto que incluyen la medida de uno o más componentes del autoconcepto físico como subescalas independientes claramente diferenciadas de otros subdominios específicos del autoconcepto. Como puede verse, en la mayoría de estos instrumentos, el autoconcepto físico se define como una dimensión del autoconcepto que a su vez se divide en apariencia física y habilidad física.

Instrumentos multidimensionales de medida del Autoconcepto	Subescalas Físicas que incorporan
Escala de Autovaloración (Fleming y Courtney, 1984)	Habilidad física Apariencia física
Perfil de Autopercepción para Niños (Harter, 1985)	Competencia atlética Apariencia física
Cuestionario de Autopercepción I (Marsh, 1988)	Habilidad física Apariencia física
Cuestionario de Autopercepción II (Marsh, 1990b)	Habilidad física Apariencia física
Cuestionario de Autopercepción III (Marsh, 1992)	Habilidad física Apariencia física
Escala Multidimensional de Autoconcepto (Bracken, 1992)	Competencia física Apariencia física Forma física Salud
Cuestionario Revisado de Autoimagen de Offer (Offer, Ostrov, Howard, y Dolan, 1992)	Imagen corporal

Fig. 4.5. Algunos de los cuestionarios multidimensionales de medida del autoconcepto que incluyen subescalas para medir componentes del autoconcepto físico.

Sin embargo, como ya hemos comentado anteriormente, los avances teóricos y empíricos en el estudio del autoconcepto llevan a considerar la necesidad de construir instrumentos multidimensionales específicos para cada una de sus dimensiones, con objeto de poder profundizar más en el estudio de la estructura de este constructo. El desarrollo de estos instrumentos ha supuesto un avance importante en el estudio del autoconcepto físico, y ha permitido definirlo como un constructo multidimensional y jerárquico en el que diferentes componentes específicos ocupan la base de la jerarquía, mientras que un autoconcepto físico global ocupa el ápice de dicha jerarquía (Fox y Corbin, 1989; Marsh y Redmayne, 1994; Marsh, Richards, Johnson, Roche, y Tremayne, 1994; Sonstroem, Speliotis, y Fava, 1992). Como

veremos a continuación, los componentes específicos varían entre los diferentes autores.

El modelo de Shavelson y cols. (1976) ha representado el punto de partida de tres importantes instrumentos de medida del autoconcepto físico: el Perfil de Autopercepción Física (PSPP) -"Physical Self Perception Profile"- (Fox, 1990; Fox y Corbin, 1989), la Escala de Autoconcepto Físico (PSC) -"Physical Self Concept Scale"- (Richards, 1987,1988); y el Cuestionario de Autodescripción Física (PSDQ) - "Physical Self Description Questionnaire"- (Marsh, Richards, Johnson, Roche y Tremayne, 1994).

Fox y Corbin (1989) desarrollaron el cuestionario PSPP en base a investigaciones teóricas y empíricas previas en el campo del autoconcepto (Harter, 1985, 1986; Shavelson y cols., 1976). El cuestionario está compuesto por cuatro subescalas que miden diferentes subdominios del autoconcepto físico (competencia en el deporte, condición física, atractivo corporal, fuerza física), y una quinta subescala que mide autovaloración física general. Fox y Corbin (1989) y Fox (1990) ofrecen evidencia empírica que apoya la estructura multidimensional jerárquica que subyace al PSPP, así como evidencia a favor de su validez de constructo. Moreno (1997) ha traducido este cuestionario al castellano y analizado sus propiedades psicométricas en una muestra de adolescentes de la Comunidad Valenciana.

Richards (1987, 1988) desarrolló el PSC para medir las diferentes dimensiones del autoconcepto físico en niños a partir de los 12 años. Para la construcción del cuestionario, el autor se basó en el modelo de Marsh/Shavelson (1985), en los cuestionarios SDQ, y en una revisión de

la literatura sobre el autoconcepto físico. El PSC está compuesto por seis subescalas que miden diferentes subdominios del autoconcepto físico (constitución corporal, apariencia física, salud, competencia física, fuerza, y orientación hacia la acción), y una séptima subescala que mide satisfacción física general. Richards (1987) analizó las propiedades psicométricas y las estructura factorial del cuestionario obteniendo resultados satisfactorios.

De la descripción del PSDQ, de su proceso de construcción, y de los estudios llevados a cabo para analizar sus propiedades psicométricas nos ocupamos ampliamente en el siguiente apartado.

4.2. El cuestionario multidimensional de autoconcepto físico PSDQ

Marsh y Redmayne (1994) describen el desarrollo de una versión preliminar del PSDQ compuesta por seis subescalas que miden diferentes componentes del autoconcepto físico: apariencia física, habilidad física, resistencia, equilibrio, flexibilidad, y fuerza. En este estudio, los autores analizaron la relación entre esos seis componentes del autoconcepto físico y cinco componentes de la forma física (resistencia, equilibrio, flexibilidad, fuerza estática, y fuerza explosiva) en una muestra de jóvenes adolescentes de 13 y 14 años. Los sujetos que formaron parte del estudio completaron el cuestionario y realizaron diferentes tests físicos (que incluían pruebas de carrera, salto, etc.) para obtener medidas de los cinco componentes de la forma física. Los coeficientes alfa de Cronbach fueron satisfactorios para todas las subescalas, presentando valores entre .84 y .92. La realización de un análisis factorial exploratorio llevó a la identificación de los seis factores

hipotetizados, y mediante el uso de análisis factorial confirmatorio se puso de manifiesto el adecuado ajuste de un modelo multidimensional jerárquico de seis factores. Además, el patrón de correlaciones entre los componentes específicos del autoconcepto físico y de la forma física ofrecieron evidencia a favor de la validez de constructo del instrumento: las subescalas resistencia, fuerza y flexibilidad presentaron correlaciones positivas y significativas con sus correspondientes componentes de la forma física; sin embargo, la subescala de equilibrio no presentó una correlación significativa con su correspondiente indicador objetivo, por lo que dicha escala no fue utilizada posteriormente en la elaboración de la versión final del PSDQ. Todos estos resultados ofrecen evidencia que apoya la estructura factorial del instrumento, y el modelo multidimensional jerárquico del autoconcepto físico que le subyace.

La versión final del PSDQ (Marsh y cols. 1994) está compuesta por 70 items que miden 9 componentes específicos del autoconcepto físico (salud, coordinación, actividad física, grasa corporal, competencia deportiva, apariencia física, fuerza, flexibilidad, y resistencia), y 2 componentes globales (autoconcepto físico global, y autoestima), lo que hace un total de 11 subescalas cuya descripción, así como el número de items que las componen, aparece reflejado en la tabla 4.6.

Physical Self Description Questionnaire (PSDQ)		
Subescala	Nº Items	Descripción de la subescala
Salud	8	No ponerse enfermo con frecuencia, recuperarse rápidamente tras una enfermedad.
Coordinación	6	Ser bueno en movimientos que requieren coordinación, ser capaz de realizar movimientos físicos con armonía.
Actividad Física	6	Ser físicamente activo, hacer muchas actividades físicas de forma regular.
Grasa Corporal	6	No tener exceso de peso, no estar demasiado gordo.
Competencia Deportiva	6	Ser bueno en los deportes, ser atlético, tener buenas habilidades deportivas.
Apariencia Física	6	Ser atractivo, tener un aspecto agradable.
Fuerza	6	Ser fuerte, tener un cuerpo fuerte y musculado.
Flexibilidad	6	Ser capaz de doblar y retorcer el cuerpo fácilmente.
Resistencia	6	Ser capaz de correr una larga distancia sin parar, no cansarse fácilmente cuando se realiza un ejercicio duro.
Autoconcepto Físico Global	6	Actitud positiva sobre el propio autoconcepto físico.
Autoestima	8	Sentimientos positivos generales sobre uno mismo.

Tabla 4.6. Descripción de las 11 subescalas que componen el PSDQ.

El modelo de Marsh/Shavelson (Marsh y Shavelson, 1985), la investigación previa llevada a cabo con el SDQ, y la versión preliminar presentada en el trabajo de Marsh y Redmayne (1994), constituyen las bases para el desarrollo del PSDQ. Este toma alguna de las subescalas del SDQ (apariencia física, y habilidad física, que ahora recibe el nombre de competencia deportiva); otras de la primera versión del PSDQ (resistencia, flexibilidad y fuerza); y se desarrollan nuevas escalas que pretenden representar las percepciones de los sujetos respecto a otros componentes de la forma física (salud, coordinación, actividad física, y grasa corporal). Estas nuevas escalas se desarrollan en base al estudio de

Marsh (1993) en el que se realizaron análisis factoriales confirmatorios con los datos de indicadores de la forma física extraídos del Estudio Australiano sobre Salud y Forma Física.

Marsh y cols. (1994) evaluaron las propiedades psicométricas del PSDQ en dos muestras de estudiantes de edades comprendidas entre 12 y 18 años. Los valores del coeficiente alfa de Cronbach fueron satisfactorios para las 11 subescalas, presentando valores entre .82 y .96. Se llevaron a cabo análisis factoriales confirmatorios para poner a prueba el ajuste del modelo hipotetizado de once factores; los valores ofrecidos por diferentes índices de ajuste pusieron de manifiesto que el modelo de once factores se ajustaba de forma adecuada a los datos, tanto en el grupo de chicos como en el de chicas, y esto se confirmaba en las dos muestras. Además, se realizaron análisis factoriales confirmatorios multigrupo en las dos muestras, para analizar la invarianza de la estructura factorial del instrumento en función del sexo. La evaluación de modelos progresivamente más restrictivos confirmó la invarianza de las saturaciones factoriales, de las varianzas y covarianzas entre factores, y de las correlaciones entre factores. Todos estos resultados indican que la estructura factorial hipotetizada se ajusta de forma adecuada a los datos, tanto en el grupo de chicos como en el de chicas, y que las soluciones factoriales son invariantes en los dos grupos.

En este mismo estudio (Marsh y cols. 1994), y en base a las respuestas de los sujetos de una de las dos muestras en los cuestionarios de autoconcepto físico PSPP y PSC, se realizaron análisis con matrices multirrasgo-multimétodo, que ofrecieron apoyo empírico a la validez convergente y discriminante de los tres instrumentos de medida.

Marsh (1996) analiza la validez de constructo del PSDQ con respecto a 23 criterios externos, entre los que se incluyen medidas de la composición corporal (índice de masa corporal, pliegues cutáneos, perímetros corporales), medidas de la participación en actividades físicas, y tests físicos de resistencia, fuerza y flexibilidad. Para cada uno de los criterios externos se hipotetizó que estaría más fuertemente correlacionado con una subescala del PSDQ en particular. Los resultados apoyan la validez convergente del PSDQ, ya que cada una de las correlaciones predichas fue estadísticamente significativa. Por otra parte, para la mayoría de los criterios externos, la correlación hipotetizada presentaba el valor más elevado frente a las correlaciones del mismo criterio con otras subescalas, lo cual apoya la validez discriminante de las respuestas del PSDQ.

Los resultados de todos estos estudios apoyan la fiabilidad y la validez del cuestionario PSDQ y lo señalan como un instrumento útil para la medida del autoconcepto físico.

A lo largo de este capítulo hemos comprobado que la investigación del autoconcepto ha llevado progresivamente a una profundización en el estudio de la estructura de los diferentes dominios que componen este constructo: autoconcepto académico, físico, social, etc. Esto ha llevado al desarrollo de cuestionarios específicos para medir estos estos dominios. Un claro ejemplo de esta evolución son los diferentes instrumentos elaborados por Marsh para medir el autoconcepto (SDQI, SDQII, y SDQIII), el autoconcepto académico (ASDQ), y el autoconcepto físico (PSDQ). Los estudios sobre este último constructo realizados con muestras de escolares y de atletas de élite en edad escolar (Marsh,

Perry, Horsely, y Roche, 1995; Marsh, Hey, Roche, y Perry, 1997), han llevado a este autor a plantear la necesidad de desarrollar un instrumento de medida del autoconcepto específico para los deportistas de élite. Esta propuesta es consistente con la idea de que los instrumentos de medida del autoconcepto más útiles, son aquéllos que miden componentes específicos que son particularmente relevantes en una determinada área de investigación. El resultado de estas reflexiones ha sido la construcción del Cuestionario de Autodescripción para Atletas de Elite (EASDQ) -"Elite Athlete Self Description Questionnaire"- (Marsh, Hey, y Johnson, en prensa), que abre una nueva vía para profundizar en el estudio del autoconcepto.

CAPITULO 5

METODO

En este capítulo se describen los aspectos de método del presente trabajo. En el apartado 1 se exponen el objetivo general y los objetivos específicos del estudio. En el siguiente apartado se presenta el procedimiento de selección de la muestra, haciendo referencia tanto a los criterios seguidos en la selección de las culturas como a los criterios seguidos en la selección de los sujetos dentro de cada cultura. En el apartado 3 se describen las características del instrumento de medida que va a ser objeto de estudio. El apartado 4 recoge el procedimiento llevado a cabo en el proceso de traducción/adaptación del instrumento de medida, el diseño de recogida de datos, y el procedimiento que se ha seguido para la recogida de datos. En el apartado 5 presentamos una descripción detallada de las características de las muestras. Hablamos de muestras ya que el estudio se ha llevado a cabo con dos muestras

procedentes de paises diferentes (Australia y España), por lo que la descripción se hace de forma separada en cada una de ellas. Finalmente, el apartado 6 presenta los análisis llevados a cabo para evaluar la equivalencia de las dos versiones del instrumento de medida.

1. OBJETIVOS

El objetivo del presente trabajo consiste en la traducción, adaptación, y análisis de la equivalencia del Physical Self Description Questionnaire (PSDQ) en su versión al castellano. Este objetivo plantea el desarrollo de dos fases: una primera fase en la que se realiza la traducción y adaptación del instrumento, y una segunda fase en la que se pone a prueba la equivalencia entre las dos versiones del instrumento (la original y la traducida), y que pretende ofrecer evidencia que garantice la idoneidad de la traducción realizada. La literatura transcultural señala la necesidad de realizar estudios que pongan a prueba la equivalencia entre las diferentes versiones de un instrumento traducido a otros idiomas (Hambleton, 1993; Hambleton y Kanjee, 1995; Van de Vijver y Hambleton, 1996; Van de Vijver y Leung, 1996). Este es un requisito imprescindible para garantizar la validez de los resultados de los estudios de comparación transcultural obtenidos en base a la aplicación de estos instrumentos. El objetivo general y los objetivos específicos que se plantean son los siguientes:

Objetivo general:

Ofrecer una versión en castellano del PSDQ que sea equivalente a la versión original en inglés del cuestionario, de tal forma que permita la comparación transcultural del autoconcepto físico entre muestras de

sujetos que pertenecen a diferentes culturas y que no comparten el mismo idioma (inglés versus castellano). Contar con instrumentos de medida equivalentes es un requisito indispensable para asegurar la validez de las comparaciones realizadas en estudios transculturales.

Objetivos específicos:

1. Llevar a cabo la traducción y adaptación del PSDQ en su versión al castellano utilizando los métodos de juicio y los consejos y recomendaciones que recoge la literatura transcultural para garantizar la adecuada traducción del instrumento.

2. Analizar la equivalencia estructural de la versión traducida al castellano del PSDQ respecto a la versión original en inglés, para comprobar si miden un mismo constructo y con las mismas dimensiones.

3. Analizar la invarianza factorial de las dos versiones del instrumento. Este objetivo específico se desglosa en varios, en función de las hipótesis de invarianza planteadas:

   3.1. Invarianza de las saturaciones factoriales.

   3.2. Invarianza de las saturaciones factoriales y de los interceptos.

2. SELECCION DE LA MUESTRA

La comparación transcultural no es el objetivo de este trabajo. Este estudio es más bien un paso previo para garantizar la validez de la comparación transcultural del autoconcepto físico medido con el PSDQ en

adolescentes australianos y españoles. La confirmación de la equivalencia de las dos versiones del instrumento permitirá dar un paso más y plantear la comparación entre las muestras procedentes de las dos culturas. Los procedimientos utilizados para la selección de las muestras y de los sujetos de cada muestra son similares a los que se siguen en cualquier estudio de comparación transcultural.

2.1. Muestreo de las culturas

En la literatura transcultural se distinguen tres tipos diferentes de muestreo para llevar a cabo la selección de las culturas que van a ser objeto de estudio (Van de Vijver y Leung, 1996): el muestreo de conveniencia, el muestreo sistemático, y el muestreo aleatorio.

En el muestreo de conveniencia (convenience sampling), la elección de las culturas se realiza por pura conveniencia. El investigador puede seleccionar la cultura simplemente por el hecho de pertenecer a ella, por tener contactos con investigadores pertenecientes a esa cultura, o elegir aquélla en la que casualmente está pasando un año sabático. En estos estudios no suele haber una hipótesis a priori sobre las diferencias entre las culturas en el constructo analizado, sino que normalmente se adopta una postura de "a ver qué pasa", y en el caso de encontrar diferencias culturales es cuando se buscan explicaciones.

En el muestreo sistemático (systematic sampling), las culturas son seleccionadas dentro de un marco teórico. La selección de las culturas se basa en que éstas representan valores diferentes en un constructo que forma parte de una teoría, y que puede ser concebido como un continuo con dos extremos. Por ejemplo, el continuo colectivismo-individualismo,

donde una cultura con fuertes valores individualistas estaría en un extremo, y una cultura con fuertes valores colectivistas estaría en el otro extremo del continuo. De este modo, se podrían plantear hipótesis sobre las diferencias a encontrar en esas dos culturas debido a su distinta situación en el continuo. Este tipo de muestreo es adecuado cuando el marco teórico que guía el estudio está plenamente desarrollado para explicar las diferencias encontradas, por lo tanto no debería utilizarse en estudios exploratorios o cuando el marco teórico no está suficientemente desarrollado. Un ejemplo de este tipo de trabajos lo encontramos en el estudio de Berry (1967). El marco teórico del que parte este trabajo es la relación entre el tipo de sociedad (agrícola-cazadora) y la influencia de ésta característica en la variable dependencia-independencia. Se hipotetiza que las sociedades agrícolas imponen mayor presión hacia la conformidad, por lo que los sujetos serán más dependientes. Por otro lado, se hipotetiza que las sociedades cazadoras animan a sus miembros a ser autónomos, y por lo tanto serán más independientes. El análisis se lleva a cabo en dos muestras, una perteneciente a una sociedad agrícola y la otra perteneciente a una sociedad cazadora, y se analizan las diferencias en la variable dependencia-independencia.

Por último, en el muestreo aleatorio (random sampling), se selecciona un gran número de culturas al azar, generalmente para evaluar una estructura universal o una teoría pan-cultural. Aunque existen pocos estudios de este tipo, podemos citar algún ejemplo como el estudio de Schwartz (1992, 1994) en el que se utilizaron muestras de 36 culturas diferentes para evaluar la estructura de los valores humanos, o el estudio de Buss et al. (1990) en el que se recogió muestra en 37 culturas para estudiar la elección de compañero.

El tipo de muestreo más utilizado en los estudios de comparación transcultural es el de conveniencia (Van de Vijver y Leung, 1996). Este es también el tipo de muestreo que se ha utilizado en este trabajo. Se utilizan dos muestras, una de adolescentes valencianos y otra de adolescentes australianos. Los motivos son claros: por un lado, este trabajo se ha realizado en una universidad de la Comunidad Valenciana, por otro lado, los contactos establecidos con el autor del cuestionario y sus colaboradores, todos ellos investigadores pertenecientes a la universidad de Western Sydney, ha permitido tener acceso a los datos de una muestra de adolescentes australianos.

2.2. Muestreo de los sujetos

El muestreo de los sujetos en los estudios transculturales puede ser de dos tipos: aleatorio (random sampling), o por emparejamiento (matched sampling). El primero hace referencia a la elección aleatoria de los sujetos que van a formar parte de la muestra. Sin embargo, hemos de tener en cuenta que para poder obtener comparaciones transculturales válidas, los sujetos de los diferentes grupos culturales deben ser equivalentes respecto a variables relevantes para el estudio. De no ser así, sería muy difícil poder discernir si las diferencias encontradas son debidas a la cultura o a las características de cada muestra. Por ejemplo, no tendría sentido comparar la comprensión lectora de un grupo de analfabetos y de un grupo de universitarios de diferentes culturas. En este caso no podríamos concluir que los miembros de una cultura presentan mayor comprensión lectora que los de la otra, ya que las dos muestras no son equivalentes en un aspecto relevante para la comparación. Un procedimiento para evitar este problema es emparejar las muestras en base a sus características

demográficas, de forma que las diferencias muestrales puedan ser excluidas como explicaciones alternativas para las diferencias culturales observadas.

En este estudio se ha llevado a cabo un muestreo por emparejamiento. Las dos muestras están compuestas por estudiantes adolescentes que presentan características similares en las siguientes variables demográficas: edad, distribución por sexo, y curso. Como veremos más adelante en el apartado de descripción de las muestras, ambas presentan una distribución similar por sexos, y los sujetos de las dos muestras pertenecen al mismo rango de edad y curso académico.

3. INSTRUMENTO DE MEDIDA

Como ya hemos comentado anteriormente, el presente trabajo se basa en el estudio del Physical Self Description Questionnaire (PSDQ) en su versión traducida al castellano. Este cuestionario está formado por 70 items que miden 9 componentes específicos del autoconcepto físico (salud, coordinación, actividad física, grasa corporal, competencia deportiva, apariencia física, fuerza, flexibilidad, y resistencia), y 2 componentes globales (autoconcepto físico global, y autoestima), lo que hace un total de 11 subescalas cuya descripción, así como el número de items que las componen, ha sido presentado en el capítulo 4 (ver tabla 4.6).

Cada ítem es un enunciado al que el sujeto debe responder en una escala de 6 puntos que va desde 1="Totalmente falso" a 6="Totalmente verdadero". En la tabla 5.1. presentamos los items que componen el

cuestionario agrupados por subescalas. El número que figura delante de cada ítem corresponde a su numeración en el cuestionario; cuando aparece el símbolo (I) delante del número del ítem, ello indica que es un ítem invertido.

Tabla 5.1. Items del cuestionario PSDQ agrupados por subescalas. (Continúa).

SALUD

(I) 1. Cuando estoy enfermo/a, me encuentro tan mal que no puedo ni levantarme de la cama.
(I) 12. Normalmente cojo todas las enfermedades (gripe, virus, resfriados, etc.) que hay por ahí.
(I) 23. Estoy enfermo/a tan a menudo que no puedo hacer todas las cosas que quisiera.
34. Casi nunca me pongo enfermo/a.
(I) 45. Me pongo enfermo/a con mucha frecuencia.
(I) 56. Cuando me pongo enfermo/a me cuesta mucho tiempo recuperarme.
(I) 67. Me pongo enfermo/a y tengo que ir al médico con más frecuencia que la mayoría de los chicos/as de mi edad.
69. Normalmente me mantengo sano/a, incluso cuando mis amigos/as se ponen enfermos.

COORDINACION

2.Me siento seguro/a realizando movimientos que requieren coordinación.
13. Me resulta fácil controlar los movimientos de mi cuerpo.
24. Soy bueno/a realizando movimientos que requieren coordinación.
35. En la mayoría de las actividades físicas, puedo realizar los movimientos con armonía.
46. Realizo con facilidad movimientos que requieren coordinación.
57. Me muevo con gracia y coordinación cuando practico deportes y actividades.

ACTIVIDAD FISICA

3. Varias veces a la semana realizo ejercicios o deportes lo suficientemente duros como para hacerme respirar fuerte.
14. Suelo hacer ejercicio o actividades que me hacen respirar fuerte.
25. Tres o cuatro veces a la semana y al menos durante media hora, hago ejercicio o actividades que me hacen respirar fuerte.
36. Hago actividades físicas (como correr, bailar, ir en bici, aerobic, gimnasia o nadar) por lo menos tres veces a la semana.
47. Practico muchos deportes, baile, gimnasia u otras actividades físicas.
58. Practico deportes, ejercicio, baile u otras actividades físicas casi todos los días.

GRASA CORPORAL

(I) 4. Estoy demasiado gordo/a.
(I) 15. Mi cintura es demasiado ancha.
(I) 26. Tengo demasiada grasa en mi cuerpo.
(I) 37. Peso demasiado.
(I) 48. Mi barriga es demasiado grande.
(I) 59. La gente piensa que estoy gordo/a.

COMPETENCIA DEPORTIVA

5. La gente piensa que soy bueno/a en los deportes.
16. Se me dan bien la mayoría de deportes.
27. La mayoría de deportes me resultan fáciles.
38. Tengo buenas habilidades deportivas.
49. Se me dan mejor los deportes que a la mayoría de mis amigos/as.
60. Juego bien en los deportes.

APARIENCIA FISICA

7. Teniendo en cuenta mi edad, soy atractivo/a.
18. Tengo una cara agradable.
29. Soy más guapo/a que la mayoría de mis amigos/as.
(I) 40. Soy feo/a.
51. Soy guapo/a.
(I) 62. Nadie piensa que soy guapo/a.

FUERZA

8. Soy una persona físicamente fuerte.
19. Tengo mucha fuerza física.
30. Soy más fuerte que la mayoría de los chicos/as de mi edad.
(I) 41. Soy débil y casi no tengo músculo.
52. Obtendría buenos resultados en una prueba de fuerza.
63. Se me da bien levantar objetos pesados.

FLEXIBILIDAD

9. Soy bastante bueno/a doblándome y retorciendo mi cuerpo.
20. Mi cuerpo es flexible.
(I) 31. Mi cuerpo es rígido y nada flexible.
42. Puedo doblar y mover bien las diversas partes de mi cuerpo en la mayoría de las direcciones.
53. Creo que tengo bastante flexibilidad para la práctica de la mayoría de los deportes.
64. Creo que obtendría buenos resultados en una prueba de flexibilidad.

RESISTENCIA

10. Puedo correr largas distancias sin parar.
21. Obtendría buenos resultados en una prueba de resistencia física.
32. Podría correr durante 5 kilómetros sin parar.
43. Creo que podría correr una distancia larga sin cansarme.
54. Puedo mantenerme físicamente activo/a durante un periodo largo de tiempo sin cansarme.
65. Se me dan bien las actividades de resistencia física, como las carreras de larga distancia, el aerobic, el ciclismo o la natación.

AUTOCONCEPTO FISICO GLOBAL

6. Físicamente, estoy satisfecho/a con el tipo de persona que soy.
17. Físicamente, me siento contento/a conmigo mismo/a.
28. Me siento satisfecho/a con mi apariencia física y con lo que puedo hacer físicamente.
39. Físicamente, me siento satisfecho/a conmigo mismo/a.
50. Me siento satisfecho/a con quien soy y con lo que puedo hacer físicamente.
61. Estoy satisfecho/a con cómo soy físicamente.

AUTOESTIMA

11. En general, la mayoría de las cosas que hago me salen bien.
(I) 22. No tengo mucho de lo que sentirme orgulloso/a.
(I) 33. Siento que mi vida no es demasiado útil.
(I) 44. En general, no valgo para nada.
55. Hago bien la mayoría de las cosas que hago.
66. En general, tengo mucho de lo que sentirme orgulloso/a.
(I) 68. En general, soy un fracaso.
(I) 70. Nada de lo que hago parece salir bien.

Las instrucciones y el formato de respuesta del cuestionario se basan en la versión comercial del SDQII (Self Description Questionnaire II) (Marsh, 1990b). En el Apéndice I presentamos la versión original en inglés del cuestionario PSDQ, así como la versión traducida al castellano.

Aunque este cuestionario fue diseñado para ser utilizado con adolescentes a partir de los 12 años, también es apropiado para su administración a adultos (Marsh et al., 1994).

4. PROCEDIMIENTO

En este apartado vamos a exponer, en primer lugar, el proceso que se llevó a cabo para la traducción y adaptación del cuestionario PSDQ. Este proceso supuso una serie de pasos de revisión, análisis, y puesta a prueba del cuestionario, que dieron como resultado la versión final en castellano que ha sido objeto de estudio en este trabajo. Posteriormente pasaremos a comentar el diseño de recogida de datos utilizado y el procedimiento seguido para la recogida de los datos.

4.1. Proceso de traducción/adaptación del instrumento de medida

En el capítulo 2 hemos hablado de tres procedimientos diferentes para llevar a cabo la adaptación de un instrumento de medida: aplicar, adaptar, o construir un instrumento completamente nuevo. El objetivo de este estudio es ofrecer un versión en castellano del PSDQ que permita la comparación del autoconcepto físico de adolescentes que hablan diferentes idiomas (inglés y castellano), por lo tanto, la opción más

adecuada en este caso es la primera. La aplicación del mismo instrumento (o de una traducción si es necesario), es también la opción que recomiendan adoptar por defecto Van de Vijver y Leung (1996), entre otras cosas, porque la recogida de información con el mismo instrumento, sin realizar ningún tipo de modificación, posibilita comparar los resultados obtenidos en una investigación con otros resultados ya publicados en la literatura.

La elección de la alternativa de aplicar una versión traducida del instrumento original, además de encajar con el objetivo de este trabajo, parece adecuada, ya que pensamos que se puede asumir la ausencia de sesgos del constructo y del método. Con respecto al sesgo del método, tanto el formato de respuesta como los tipos de estímulos presentados son familiares en el nuevo contexto cultural en el que va a ser aplicado el instrumento, por lo que no hay ningún motivo que induzca a pensar en la necesidad de introducir modificaciones en el formato del cuestionario para evitar este sesgo. Respecto al sesgo del constructo, se parte de la base de que el autoconcepto físico presenta las mismas características en los dos grupos. No obstante, la metodología utilizada para analizar la equivalencia de las dos versiones puede ofrecernos información valiosa que ponga de manifiesto diferencias entre los dos grupos analizados respecto al constructo medido, en el caso de que éstas realmente existan.

La traducción fue llevada a cabo por un equipo de personas, entre las que se encontraban dos traductoras profesionales, todas ellas con un nivel adecuado de competencia en los dos idiomas. La profesión y procedencia de éstas personas hace que cada una de ellas aportara al equipo de investigación alguno de los aspectos importantes y necesarios

para garantizar la traducción adecuada del cuestionario: conocimiento de la cultura española, conocimiento de la cultura anglosajona, conocimiento de la materia sobre la que trata el cuestionario, y conocimientos sobre la construcción de tests y la elaboración de items.

Se utilizaron métodos de juicio para llevar a cabo la traducción. En concreto, se recurrió al procedimiento de traducción hacia atrás o back translation. Los pasos que se siguieron se exponen a continuación. El término cultura objetivo hace referencia a aquella en la que va a ser traducido y adaptado el instrumento de medida (en este caso la española).

1° Un grupo formado por dos personas pertenecientes a la cultura objetivo, y con conocimientos sobre la materia y sobre la construcción de tests y elaboración de items, tradujo el cuestionario PSDQ del inglés al castellano.

2° De forma paralela pero independiente, una traductora profesional perteneciente también a la cultura objetivo, realizó otra traducción del cuestionario del inglés al castellano.

3° Tras poner en común y contrastar las dos traducciones paralelas en sucesivas reuniones y sesiones de trabajo, se obtuvo una primera versión en castellano del cuestionario.

4° Esta primera versión en castellano fue entregada a otra traductora profesional perteneciente a la cultura anglosajona, para que la tradujera de nuevo al inglés.

5° Se comparó la versión original del instrumento con la traducción inversa en el mismo idioma, llevando a cabo análisis racionales sobre su equivalencia. Las personas implicadas en el proceso de traducción se reunieron y analizaron conjuntamente la equivalencia de los items uno a uno. Se detectaron algunos items cuya equivalencia no era totalmente satisfactoria a juicio de los traductores. En estos casos, se revisó la versión en castellano del ítem y se introdujeron las modificaciones oportunas para mejorar dicha equivalencia. De este modo se obtuvo una segunda versión en castellano del cuestionario.

6° Se llevó a cabo un primer pase piloto para poner a prueba la traducción realizada. La versión en castellano del PSDQ obtenida hasta el momento se basaba en el trabajo de un equipo de expertos especialistas en la materia y en tareas de traducción, sin embargo, considerando que el cuestionario iba a ser administrado a adolescentes de 12 a 16 años, pensamos que sería valioso contar con la colaboración de sujetos con estas características para evaluar la adecuación del cuestionario. Se partió de la base de que si el cuestionario no creaba problemas de comprensión en los grupos de menor edad, tampoco los crearía en los de mayor edad. De este modo, se administró el cuestionario a un total de 27 adolescentes de edades comprendidas entre 12 y 13 años que cursaban 1° de E.S.O. La administración tuvo lugar en el aula, dentro del horario de clases. Tras informar brevemente a los estudiantes del propósito de la investigación, se les pidió su colaboración, recalcando que gracias a ésta sería posible elaborar un instrumento adecuado que sería administrado posteriormente a un millar de estudiantes de la Comunidad Valenciana. Se leyeron las instrucciones en voz alta y se les pidió que contestaran el cuestionario. Una vez que todos hubieron finalizado, se estableció un diálogo abierto en el que se preguntó a los

adolescentes sobre la facilidad de comprensión de las instrucciones, del formato de respuesta y de los items. Se detectaron algunos items que por el vocabulario utilizado no les eran fácilmente comprensibles, y en esos casos se les pidió que propusieran formas alternativas de decir lo mismo pero que fueran más comprensibles para adolescentes de su misma edad. Por ejemplo, en los items 3, 14, y 25, las expresiones "breathe hard" y "huff and puff" habían sido traducidas como "jadear"; esta palabra creaba problemas de comprensión, por lo que fue sustituida por una de las expresiones que propusieron los estudiantes ("respirar fuerte"). Otro ejemplo lo encontramos en el ítem 9, que había sido traducido como "Soy bastante bueno/a doblándome y contorsionándome"; este último verbo también creaba problemas de comprensión, y en base a las alternativas que dieron los estudiantes para sustituirlo (ser flexible, doblarse, ser de goma, girarse, retorcer el cuerpo), el ítem se tradujo finalmente como "Soy bastante bueno/a doblándome y retorciendo mi cuerpo". En el ítem 12, la expresión "whatever illness is going around" había sido traducida como "todas las enfermedades que pululan por ahí", y en base a las dificultades de comprensión detectadas, esta expresión fue modificada, y traducida como "todas las enfermedades que hay por ahí". Las modificaciones introducidas tras este pase piloto dieron como resultado una tercera versión en castellano del PSDQ.

7° Se realizó un segundo pase piloto para comprobar si se habían resuelto los problemas de comprensión detectados en algunos items en el pase piloto anterior. En esta ocasión se administró el cuestionario a un total de 23 adolescentes de edades comprendidas entre 12 y 13 años, y que cursaban también 1° de E.S.O. El procedimiento de administración fue idéntico al realizado en el caso anterior. En esta ocasión el diálogo

con los estudiantes llevó a concluir que esa versión del cuestionario no presentaba problemas de comprensión, por lo que se dio por terminado el proceso de traducción/adaptación del cuestionario. Esta versión definitiva es objeto de análisis en el presente trabajo.

Por último, señalar que para asegurar al máximo la fidelidad de la versión en castellano respecto a la original en inglés, se respetó incluso el formato del cuestionario, tanto en las instrucciones, como en la distribución de las columnas de items (ver apéndice 1).

4.2. Diseño de recogida de datos

En el capítulo 3 hemos comentado los tres diseños estadísticos de recogida de datos que aparecen en la literatura transcultural en función de la versión del instrumento administrada (versión original, traducción inversa o versión adaptada), y de las características de la muestra de sujetos a los que se administra el instrumento de medida (monolingües o bilingües).

En base a los objetivos planteados en este estudio, el diseño más adecuado es el de aplicación de la versión original a sujetos monolingües en el idioma fuente (inglés), y de la versión adaptada a sujetos monolingües en el idioma objetivo (español). De este modo, y como ya hemos comentado anteriormente, se cuenta con dos muestras: una muestra compuesta por adolescentes australianos a los que se les ha administrado la versión original en inglés del PSDQ, y una muestra de adolescentes de la Comunidad Valenciana a los que se ha administrado la versión adaptada al castellano de dicho cuestionario.

Este diseño permite analizar la equivalencia de las dos versiones del instrumento de medida, y su principal ventaja es que al utilizar muestras de las poblaciones fuente y objetivo, los resultados sobre la equivalencia de las dos versiones del instrumento pueden generalizarse a las poblaciones de interés (teniendo siempre en cuenta que las muestras sean representativas de las poblaciones respectivas). Sin embargo, esta generalización no es posible cuando se utiliza cualquiera de los otros dos diseños de recogida de datos comentados en el capítulo 3 (aplicación de las versiones original y adaptada del instrumento a sujetos bilingües; y aplicación de la versión original y de la traducción inversa a sujetos monolingües en el idioma fuente).

La aplicación de procedimientos estadísticos en base a los datos recogidos con este diseño permite detectar los items que funcionan diferencialmente en los dos grupos culturales. Sin embargo, como ya hemos comentado anteriormente, existe una limitación, ya que ni el diseño ni los procedimientos utilizados ofrecen información para determinar si tales diferencias se deben a una adaptación deficiente del instrumento, o a diferencias reales entre las culturas.

4.3. Procedimiento de recogida de datos

La muestra de adolescentes australianos fue recogida y facilitada por miembros del equipo de investigación que ha desarrollado el cuestionario PSDQ (Marsh, Richards, Johnson, Roche, y Tremayne, 1994), y que pertenece a la Universidad de Western Sydney. Estos se pusieron en contacto con los dos centros de educación secundaria seleccionados para solicitar su colaboración. Los centros enviaron a los padres de los alumnos una hoja en la que se les explicaban los objetivos de la

investigación, y se les pedía que autorizaran a su hijo/a a participar en el estudio. Unicamente cumplimentaron el cuestionario aquellos estudiantes que habían entregado la autorización debidamente cumplimentada y firmada por sus padres y por ellos mismos. Miembros del equipo de investigación se reunieron con los profesores de los diferentes centros. Se les pidió que cumplimentaran ellos mismo el cuestionario para familiarizarse con él, y posteriormente se les dieron instrucciones sobre cómo administrarlo. Por lo tanto, la administración fue llevada a cabo por los propios profesores de cada centro, en clases que nunca superaban los 30 alumnos.

Para asegurar la estandarización del procedimiento de aplicación se pidió a los profesores que siguieran las instrucciones siguientes:

1° Entregar una copia del cuestionario a cada alumno, y asegurarse de que todos tienen bolígrafo para poder cumplimentarlo. El cuestionario no debe ser rellenado con lápiz.

2° Pedir a los estudiantes que rellenen la información que figura al principio del cuestionario (nombre, edad, sexo, centro, y curso).

3° Informar a los estudiantes que sus respuestas serán tratadas con absoluta confidencialidad y que únicamente van a ser utilizadas para dicha investigación. Solicitar su sinceridad a la hora de cumplimentar el cuestionario.

4° Leer en voz alta las instrucciones que figuran impresas en la primera hoja del cuestionario. Pedir a los estudiantes que escuchen o que sigan al mismo tiempo la lectura en su hoja. Hacer una pausa

después del ejemplo número 3 para que los estudiantes puedan marcar su respuesta.

5° No suele haber problemas a la hora de comprender el procedimiento de respuesta, sin embargo, es importante asegurarse de que todos los estudiantes han entendido cómo marcar la respuesta y cómo corregirla en caso de equivocación. Si surge alguna duda al respecto, será necesario volver a explicarlo en voz alta. Antes de comenzar a rellenar el cuestionario deben responderse todas las preguntas que planteen los alumnos.

6° Cuando los alumnos estén preparados para comenzar, decir en voz alta: "Ahora podéis empezar a rellenar el cuestionario". Una vez se ha comenzado, no se debe hablar. Además, es importante interrumpir cualquier comentario o consulta entre los estudiantes, y cualquier vocalización en voz alta de éstos, ya sea deliberada o inconscientemente.

7° Si algún estudiante levanta la mano, acercarse hasta su mesa para responder su pregunta. Si el problema es que no entiende alguna palabra o expresión, parafrasear dicha palabra o expresión sin cambiar el significado de la frase. Pedirle al estudiante que responda lo mejor que pueda.

La muestra española fue recogida por la persona que ha desarrollado el presente trabajo, con la inestimable ayuda de un miembro del equipo de investigación del que forma parte, ambas pertenecientes a la Universitat de València. El objetivo era conseguir una muestra representativa de adolescentes de la Comunidad

Valenciana. Para ello se hizo un sorteo aleatorio entre todos los Servicios Psicopedagógicos Escolares (SPEs) y los Gabinetes Psicopedagógicos Municipales (GPMs) de la Comunidad Valenciana. El número de SPEs y GPMs elegidos fue proporcional al número total de ellos en cada provincia. Se contactó por teléfono con los psicólogos/as de los servicios y gabinetes psicopedagógicos seleccionados y se concertaron citas con cada uno de ellos para explicarles personalmente el objetivo y contenido de nuestra investigación. En esta reunión se pedía a los psicólogos su colaboración para facilitarnos el acceso a los centros de educación secundaria y/o bachillerato en los que desarrollaban su actividad profesional, y se les entregaba un informe sobre el proyecto de investigación a realizar para que lo presentaran a los directores o jefes de estudios de dichos centros. Solamente en una ocasión fue necesario concertar una cita posterior con el director de un centro. Lo habitual fue que los psicólogos sirvieran de puente de enlace, siendo ellos los que nos concertaban la cita para realizar el pase en cada centro.

Unicamente en los casos en los que el director del centro lo consideró necesario, y aprovechando una reunión ordinaria de la APA, se informó a los padres de los alumnos del proyecto de investigación que se estaba llevando a cabo, y tras explicarles los objetivos de dicha investigación, se solicitó su autorización para realizar el pase de cuestionarios.

La administración de los cuestionarios tuvo lugar en las aulas de los diferentes centros, dentro del horario normal de clases, y con un número máximo de 35 alumnos por aula. La aplicación fue llevada a cabo por miembros de nuestro equipo de investigación siguiendo el mismo procedimiento estandarizado que hemos expuesto anteriormente.

En estudios como éste, en los que se pretende analizar la equivalencia entre dos versiones del mismo cuestionario en diferentes idiomas, es habitual controlar la variable "primera lengua del sujeto" para evitar que ésta contamine los resultados. Se parte de la base de que aquellas personas que se encuentran ocasionalmente en el país por motivos de trabajo o de estudios, pero cuya primera lengua no es la del cuestionario que se les administra, pueden presentar deficiencias lingüísticas o problemas de comprensión que afecten a los resultados (Ellis, 1989). Por ejemplo, Ellis llevó a cabo un estudio para analizar la equivalencia de la traducción al alemán de un test de inteligencia americano, el Career Ability Placement Survey (CAPS; Knapp y Knapp, 1976), y de forma paralela, analizar la equivalencia de la traducción al inglés de un test de inteligencia alemán, el WILDE-Intelligenz-Test (WIT; Jäger, 1963; Jäger y Althoff, 1983). De la muestra de sujetos alemanes, obtenida a través del servicio nacional de empleo y que estaba compuesta por jóvenes en busca de su primer trabajo, se eliminó a aquellos jóvenes que no habían nacido en Alemania. Del mismo modo, de la muestra de sujetos americanos, compuesta por estudiantes universitarios, se eliminó a los estudiantes extranjeros y a los sujetos americanos cuya primera lengua no era el inglés.

En vista de todo ello, nos planteamos la necesidad de controlar que la muestra recogida estuviera compuesta únicamente por adolescentes españoles, y cuya primera lengua fuera el castellano. El control de la lengua es todavía más relevante en nuestro caso, si tenemos en cuenta que la Comunidad Valenciana es una comunidad bilingüe. Por ello, el pase del cuestionario se realizó únicamente en aquellos centros con línea de escolarización en castellano, y en los casos en los que el centro contaba con varias líneas de escolarización (Castellano, Valenciano y/o

Inmersió), el cuestionario fue administrado únicamente a los alumnos que recibían la enseñanza en castellano. Además, junto con el PSDQ, se entregaba a los estudiantes una encuesta socio-lingüística, donde se les preguntaba entre otras cosas, el lugar de nacimiento y el idioma habitual que utilizaban en casa. De este modo se detectaron dos niños croatas y una niña rusa que fueron eliminados de la muestra.

Este intento de control de la variable lingüística dio también como resultado una reducción del ámbito geográfico de procedencia de la muestra. Curiosamente, los psicólogos de los SPEs y GPMs de la provincia de Castellón que habían sido seleccionados mediante el sorteo aleatorio, trabajaban en colegios que únicamente tenían línea de escolarización en valenciano. Por lo tanto, dichos centros fueron eliminados como candidatos para realizar la administración del cuestionario por no adecuarse a los requisitos de la muestra para este estudio. Esto mismo sucedió en muchos de los SPEs y GPMs seleccionados de la provincia de Alicante. En la provincia de Valencia, el número de servicios y gabinetes psicopedagógicos con respecto a las otras dos provincias es mucho mayor. En base a esta proporción, también fue mayor el número de SPEs y GPMs seleccionados en el sorteo aleatorio pertenecientes a la provincia de Valencia. Se desestimaron aquéllos que no permitían acceder a centros con línea de escolarización en castellano, sin embargo, también fueron muchos los que reunían los requisitos exigidos a la muestra. Todas estas circunstancias, y el hecho de que la muestra recogida en la provincia de Valencia ya alcanzara e incluso superara el tamaño de la muestra australiana, nos llevaron a reducir el ámbito geográfico de procedencia de la muestra, a escolares adolescentes de la provincia de Valencia.

5. DESCRIPCION DE LAS MUESTRAS

Este estudio se ha realizado en base a los datos de 1972 adolescentes, divididos en dos grupos que representan la muestra australiana (N=986) y la muestra española (N=986). En este apartado presentamos una descripción de las dos muestras, que como veremos a continuación, son similares entre sí respecto a las principales variables sociodemográficas: edad, curso, y distribución por sexo.

5.1. Muestra Australiana

La muestra está compuesta por un total de 986 adolescentes australianos de edades comprendidas entre los 12 y los 16 años. La media de edad es de 13.5, presentando una desviación típica de 1.11. El 54.5 % de los sujetos de la muestra son varones, y el 45.5 % son mujeres. Todos ellos son estudiantes de educación secundaria, y pertenecen a dos centros públicos situados en la zona metropolitana suroeste de Sydney. Respecto al nivel socioeconómico, en ambos centros se observa una gran diversidad, desde sujetos con un nivel bajo y medio-bajo, a sujetos con un nivel socioeconómico medio-alto.

Los alumnos pertenecen a los cursos 7° a 10° del sistema educativo australiano. La correspondencia de estos niveles con los del nuevo sistema de enseñanza secundaria español, y de éstos con los del sistema antiguo de enseñanza en España, es la siguiente:

Niveles del Sistema Educativo Australiano	Niveles del Actual Sistema Educativo Español	Niveles del Antiguo Sistema Educativo Español
Year 7	1° E.S.O.	7° E.G.B.
Year 8	2° E.S.O.	8° E.G.B.
Year 9	3° E.S.O.	1° B.U.P.
Year 10	4° E.S.O.	2° B.U.P.

El Westfields Sports High School (WSHS) es uno de los dos centros en los que se administró el cuestionario, y resulta ser también uno de los centros de educación secundaria más prestigiosos en Australia. Dicho prestigio es debido a que entre sus alumnos se encuentran jóvenes atletas de élite procedentes de todo el país que compaginan sus estudios con el desarrollo de un programa de entrenamiento para deportistas de élite. Sin embargo, estos estudiantes solamente representan un pequeño porcentaje del total de los alumnos, ya que este centro también admite a estudiantes de la región, que no participan en el programa de entrenamiento, y que por lo tanto reciben una enseñanza secundaria normal. Todos los sujetos de la muestra que forman parte de este centro, pertenecen al grupo de estudiantes que no son deportistas de élite.

5.2. Muestra Española

Se recogieron datos de un total de 1044 adolescentes de la provincia de Valencia. Para homogeneizar las dos muestras, se eliminaron de la española, aquellos sujetos cuya edad no entraba dentro del rango de edades representado en la muestra australiana, eliminando un total de 14 sujetos cuya edad era de 11, 17 o 18 años. De la muestra resultante, compuesta por 1030 adolescentes de la provincia de Valencia, se eliminaron al azar otros 44 casos, todos ellos mujeres, hasta igualar el tamaño de la muestra australiana, y con el objeto de conseguir una distribución similar por sexos en las dos muestras.

La muestra en la que se basan los análisis presentados en este trabajo, está compuesta por un total de 986 adolescentes españoles de edades comprendidas entre los 12 y los 16 años. La media de edad es de 13.3, presentando una desviación típica de 1.07. El 50.6 % de los sujetos

de la muestra son varones, y el 49.4 % son mujeres. Todos ellos son estudiantes de educación secundaria, y pertenecen a catorce centros públicos de diez poblaciones diferentes de la provincia de Valencia.

En la hoja de variables sociolingüísticas se preguntaba a los sujetos la ocupación del padre y de la madre. La tabla 5.2. recoge la distribución de estas variables y nos hace entrever que el nivel socioeconómico de los sujetos que componen la muestra española también presenta una amplia diversidad.

	Ocupación del Padre		Ocupación de la Madre	
	Frecuencia	Porcentaje	Frecuencia	Porcentaje
Directivo	19	1,9	1	0,1
Profesional con titulación superior	36	3,7	12	1,2
Profesional con titulación media	41	4,2	43	4,4
Profesional con F.P. de 1° o 2° grado	187	19,0	67	6,8
Profesional no cualificado	590	59,8	217	22,0
Autónomo	53	5,4	28	2,8
Jubilado / Pensionista	7	0,7	--	--
Parado	4	0,4	--	--
Ama de casa	1	0,1	594	60,2
Dato faltante	48	4,9	24	2,4
Total	986	100,0	986	100,0

Tabla. 5.2. Distribución de frecuencias de las variables "ocupación del padre" y "ocupación de la madre" de los sujetos de la muestra española.

Hemos visto que las dos muestras utilizadas en este estudio son similares entre sí respecto a las principales variables sociodemográficas: distribución por sexo, y rango de edad y curso. No obstante, se llevaron

a cabo diferentes pruebas estadísticas para confirmar la igualdad en las dos muestras de la proporción de varones y mujeres, y de la media y varianza de edad de los sujetos.

Se realizó una prueba estadística para comprobar la hipótesis de la igualdad de las proporciones de varones y de mujeres en las dos muestras independientes (Amón, 1987, pag. 316), confirmando que la distribución por sexos es similar en las dos muestras ($z=1.74$, $p < 0.05$).

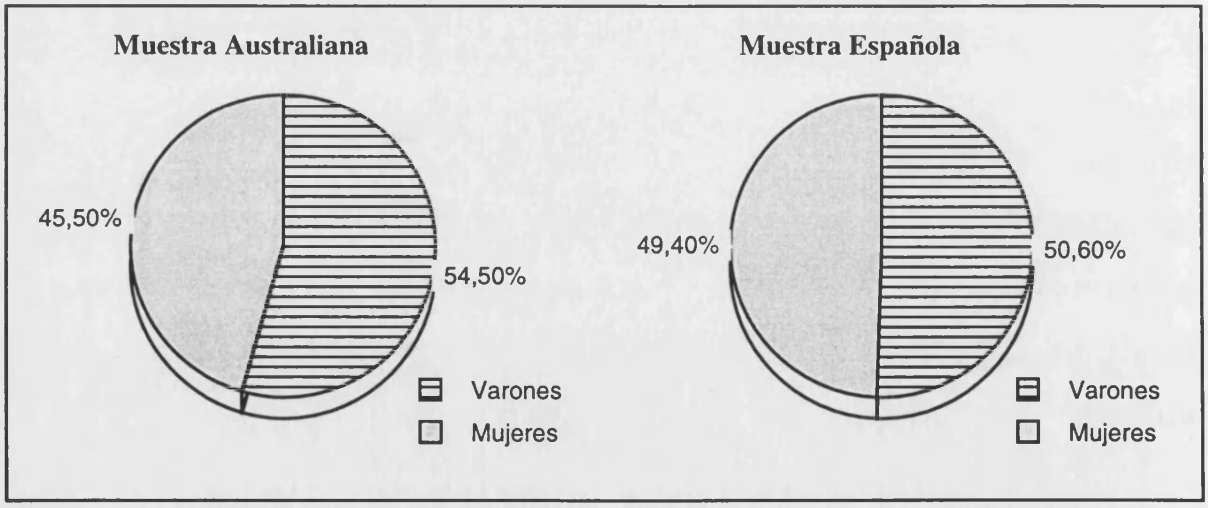
Respecto a la edad, los valores obtenidos en las muestras australiana (media=13.5, D.T.=1.11) y española (media=13.3, D.T.=1.07) parecen indicar que la edad presenta una distribución homogénea en las dos muestras, tanto en tendencia central como en variabilidad. Para confirmarlo, se realizaron también pruebas estadísticas de comprobación de hipótesis acerca de la igualdad de las medias (Amón, 1987, pag. 294) y de las varianzas de la edad de los sujetos (Amón, 1987, pag. 306) en las dos muestras independientes. En este caso, tanto la prueba de comparación de medias ($z=5.10$), como la prueba de comparación de varianzas ($F=1.08$) indican la no igualdad de estos parámetros en las dos muestras ($p<0.01$ y $p<0.05$, respectivamente). Sin embargo, estos resultados pueden explicarse en base a la potencia de las pruebas utilizadas y al gran tamaño de la muestra.

En las páginas siguientes presentamos los estadísticos descriptivos, las tablas de distribución de frecuencias y las gráficas correspondientes a las variables demográficas sexo, edad y curso, para las dos muestras.

Variable Sexo:

	Muestra Australiana		Muestra Española	
Sexo	Frecuencia	Porcentaje	Frecuencia	Porcentaje
Mujeres	448	45.4	487	49.4
Varones	537	54.5	499	50.6
Dato faltante	1	0.1	--	--
Total	986	100.0	986	100.0

Tabla. 5.3. Distribución de frecuencias de los sujetos de las dos muestras en la variable sexo.

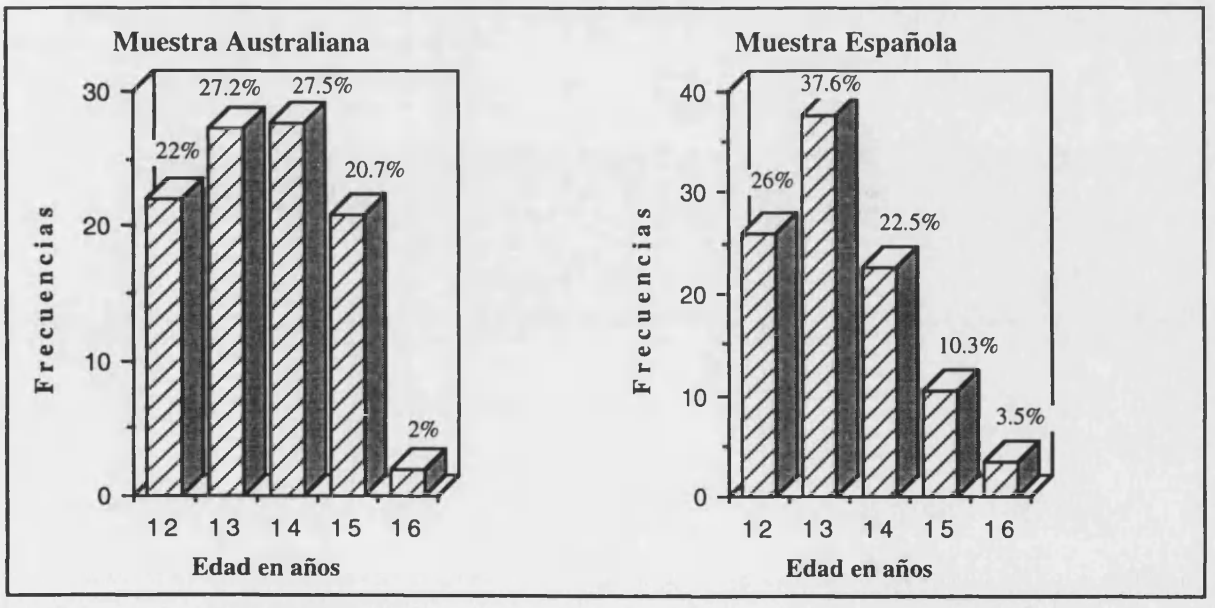


Gráfica. 5.1. Distribución de frecuencias de los sujetos de las dos muestras en la variable sexo.

Variable Edad:

	Muestra Australiana (Media = 13.5, D.T. = 1.11)		Muestra Española (Media = 13.3, D.T. = 1.07)	
Edad	Frecuencia	Porcentaje	Frecuencia	Porcentaje
12 años	217	22.0	256	26.0
13 años	268	27.2	371	37.6
14 años	271	27.5	222	22.5
15 años	204	20.7	102	10.3
16 años	20	2.0	35	3.5
Dato faltante	6	0.6	--	--
Total	986	100.0	986	100.0

Tabla. 5.4. Distribución de frecuencias de los sujetos de las dos muestras en la variable edad.

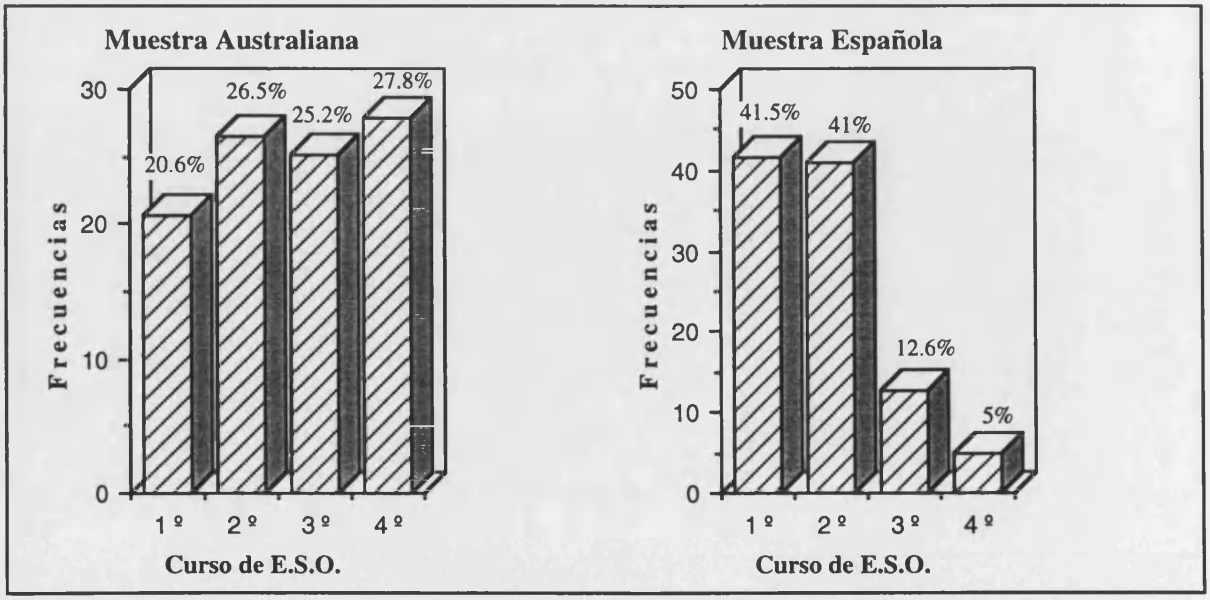


Gráfica. 5.2. Distribución de frecuencias de los sujetos de las dos muestras en la variable edad.

Variable Curso Académico:

	Muestra Australiana		Muestra Española	
Curso	Frecuencia	Porcentaje	Frecuencia	Porcentaje
1° E.S.O.	203	20.6	409	41.5
2° E.S.O.	261	26.5	404	41.0
3° E.S.O.	248	25.2	124	12.6
4° E.S.O.	274	27.8	49	5.0
Total	986	100.0	986	100.0

Tabla. 5.5. Distribución de frecuencias de los sujetos de las dos muestras en la variable curso.



Gráfica. 5.3. Distribución de frecuencias de los sujetos de las dos muestras en la variable curso.

6. ANALISIS DE DATOS

Los análisis que se llevaron a cabo, y cuyos resultados presentamos en el capítulo 6, son los siguientes:

1° Análisis de los estadísticos descriptivos de las 11 subescalas del PSDQ, de forma separada en cada una de las dos muestras:

* Media y desviación típica de cada subescala.
* Coeficiente de fiabilidad de cada subescala.
* Correlaciones entre las subescalas.

2° Análisis factorial confirmatorio con el programa LISREL 8 (Jöreskog y Sörbom, 1993) para poner a prueba la equivalencia estructural de las dos versiones del instrumento. Esto implica llevar a cabo un análisis multimuestra. Para ello es necesario calcular la matriz de covarianzas entre las variables observables (S_g) dentro de cada grupo. Se asume que hay un total de 11 factores latentes correlacionados que subyacen a las 70 variables observables y que dan cuenta de las covarianzas observadas entre ellas. Por lo tanto, se debe poner a prueba el ajuste de un modelo de 11 factores para cada matriz S_g. Esto implica definir las matrices Λ_g, Φ_g y Θ_g, es decir, determinar los parámetros libre y fijos en cada una de ellas, y las relaciones entre ellos. En este modelo, las restricciones únicamente afectan a la estructura, es decir, se pone a prueba si el patrón de relaciones entre variables observables, factores latentes y unicidades es el mismo en las dos muestras analizadas. Sin embargo, los valores de los parámetros de cada una de estas matrices son estimados libremente, es decir, no se impone ninguna constricción respecto a la invarianza de los valores de los parámetros a través de las dos muestras. Este es el modelo base que

servirá para evaluar el ajuste de los modelos más restrictivos que serán puestos a prueba posteriormente.

3° Análisis factoriales confirmatorios con el programa LISREL 8 (Jöreskog y Sörbom, 1993) para poner a prueba las diferentes hipótesis de invarianza:

3.1. Invarianza total de las saturaciones factoriales.

3.2. De no confirmarse la hipótesis anterior se pondrá a prueba un modelo menos restrictivo de invarianza parcial de las saturaciones.

3.3. Invarianza total o parcial de las saturaciones factoriales (en función de los resultados obtenidos en el análisis de las hipótesis anteriores) e invarianza total de los interceptos.

3.4. De no confirmarse la hipótesis anterior se pondrá a prueba un modelo menos restrictivo de invarianza total o parcial de las saturaciones factoriales y de invarianza parcial de los interceptos.

CAPITULO 6

RESULTADOS

En este capítulo presentamos los resultados de los análisis realizados para poner a prueba la equivalencia de las dos versiones (original en inglés y traducida al castellano) del PSDQ. En el primer apartado presentamos, de forma separada para cada una de las muestras, los estadísticos descriptivos y los coeficientes de fiabilidad de las 11 subescalas del PSDQ. En el segundo apartado figuran los resultados del ajuste de los sucesivos modelos anidados que han sido puestos a prueba para analizar la equivalencia estructural de las dos versiones del intrumento, y la invarianza de las saturaciones factoriales y de los interceptos de cada uno de los items que lo componen. Finalmente, en el apartado 3 aparecen los resultados relativos al análisis del funcionamiento diferencial de los items, indicando aquellos items que han presentado funcionamiento diferencial y el tipo de FDI presentado.

1. ESTADISTICOS DESCRIPTIVOS DE LAS SUBESCALAS DEL PSDQ

En la tabla 6.1 aparecen los estadísticos descriptivos (media y desviación típica) y los coeficientes de fiabilidad (Alfa de Cronbach) de las subescalas del PSDQ para cada una de las dos muestras. Como puede observarse en la tabla, la consistencia interna de todas las subescalas es satisfactoria, presentando valores entre .82 y .94 en la muestra australiana y entre .79 y .93 en la muestra española.

	Muestra Australiana			Muestra Española		
Subescala	Media	D.T.	Alfa	Media	D.T.	Alfa
Salud	4.73	.97	.82	4.68	.86	.79
Coordinación	4.34	1.04	.84	4.39	.94	.83
Actividad Física	4.22	1.29	.86	4.19	1.30	.89
Grasa Corporal	4.61	1.46	.94	4.58	1.31	.93
Competencia Deportiva	4.18	1.29	.93	4.19	1.14	.91
Apariencia Física	3.73	1.24	.89	3.78	1.08	.87
Fuerza	4.18	1.15	.89	3.84	1.14	.89
Flexibilidad	4.19	1.13	.85	3.86	1.10	.86
Resistencia	3.64	1.37	.90	3.78	1.25	.89
Autoconcepto Físico Global	4.43	1.28	.93	4.61	1.19	.92
Autoestima	4.73	.99	.86	4.51	.92	.84

Tabla. 6.1. Media, desviación típica y coeficiente Alfa de Cronbach de las 11 subescalas del PSDQ en las muestras australiana y española.

En las tablas 6.2. y 6.3. aparecen las correlaciones entre las subescalas del PSDQ para las muestras australiana y española respectivamente. Las intercorrelaciones de las subescalas en la muestra austaliana oscilaron entre .10 y .71. En la muestra española, las intercorrelaciones entre las subescalas oscilaron entre .01 y .66.

	SALU	COOR	ACTF	GRAS	COMP	APAR	FUER	FLEX	RESI	AFG	AEST
SALU	1.00										
COOR	.21	1.00									
ACTF	.10	.60	1.00								
GRAS	.15	.35	.23	1.00							
COMP	.16	.71	.61	.31	1.00						
APAR	.16	.47	.32	.43	.50	1.00					
FUER	.22	.49	.40	.10	.59	.49	1.00				
FLEX	.21	.69	.49	.32	.56	.42	.41	1.00			
RESI	.18	.66	.63	.45	.71	.44	.45	.60	1.00		
AFG	.17	.60	.44	.54	.59	.65	.47	.48	.56	1.00	
AEST	.27	.56	.40	.44	.51	.63	.47	.48	.48	.71	1.00

Tabla. 6.2. Correlaciones entre las subescalas del PSDQ. Muestra australiana.
Todas las correlaciones son significativas, $p < 0.01$.

	SALU	COOR	ACTF	GRAS	COMP	APAR	FUER	FLEX	RESI	AFG	AEST
SALU	1.00										
COOR	.18	1.00									
ACTF	.02 ns	.47	1.00								
GRAS	.16	.30	.18	1.00							
COMP	.11	.66	.55	.31	1.00						
APAR	.17	.36	.24	.32	.39	1.00					
FUER	.11	.46	.42	.01 ns	.55	.38	1.00				
FLEX	.08*	.45	.30	.27	.40	.29	.24	1.00			
RESI	.12	.57	.60	.35	.65	.34	.48	.39	1.00		
AFG	.21	.44	.31	.43	.48	.59	.39	.24	.42	1.00	
AEST	.31	.43	.22	.29	.36	.54	.34	.23	.31	.65	1.00

Tabla. 6.3. Correlaciones entre las subescalas del PSDQ. Muestra española.
ns = correlación no significativa.
* = Correlación significativa, $p < 0.05$.
Todas las demás correlaciones son significativas, $p < 0.01$.

Nota: clave del nombre de las subescalas.

SALU = Salud
COOR = Coordinación
ACTF = Actividad física
GRAS = Grasa corporal
COMP = Competencia deportiva
APAR = Apariencia física
FUER = Fuerza
FLEX = Flexibilidad
RESI = Resistencia
AFG = Autoconcepto físico global
AEST = Autoestima

2. ANALISIS DEL AJUSTE DE LOS DIFERENTES MODELOS MEDIANTE AFC

En este apartado presentamos el proceso seguido en la evaluación del ajuste de los diferentes modelos anidados que han sido puestos a prueba, además de los resultados obtenidos en dichos análisis. El ajuste de los modelos se evalúa en base a diferentes índices de bondad de ajuste, cuya descripción e interpretación aparece en la tabla 6.4. La elección de estos índices se basa en su uso generalizado en estudios de características similares en los que también se pretende poner a prueba el ajuste de diferentes modelos (Marsh, Balla, y McDonald, 1988).

Indice	Descripcción	Interpretación
χ^2	Chi-cuadrado	
RMSEA	Root Mean Square Error of Aproximation	RMSEA = 0 --> ajuste perfecto RMSEA < 0.05 --> ajuste satisfactorio RMSEA < 0.08 --> ajuste aceptable RMSEA > 0.1 --> ajuste inadecuado Nota: el valor p asociado indica la probabilidad de que RMSEA < 0.05
RMSRS	Root Mean Square Residual Standardized	RMSRS = 0 --> ajuste perfecto RMSRS < 0.05 --> ajuste satisfactorio RMSRS < 0.08 --> ajuste aceptable RMSRS > 0.1 --> ajuste inadecuado
GFI	Goodness of Fit Index	GFI > .90 --> ajuste satisfactorio
CFI	Comparative Fit Index	CFI > .90 --> ajuste satisfactorio

Tabla 6.4. Descripción de los índices de bondad de ajuste utilizados en este trabajo para analizar el ajuste de los diferentes modelos evaluados.

Partiendo de un modelo base que plantea la equivalencia de la estructura factorial de las dos versiones del PSDQ, se evaluó el ajuste de sucesivos modelos anidados que imponían mayor o menor número de restricciones. La tabla 6.5. presenta los índices de bondad de ajuste de los principales modelos evaluados, que representan las hipótesis de invarianza planteadas en el capítulo 5.

Modelo	χ^2	gl	χ^2_{dif}	gl_{dif}	RMSEA	p-value	RMSRS	GFI	CFI
Modelo 1	13449.05*	4580			0.031	1.00	0.053	0.82	0.90
Modelo 2	14124.22*	4639			0.032	1.00	0.054	0.81	0.89
(2/1)			675.2*	59					
Modelo 3	13511.94*	4625			0.031	1.00	0.052	0.82	0.90
(3/1)			62.9 ns	45					
Modelo 4	14911.39*	4695			0.033	1.00	0.051	0.81	0.88
(4/3)			1399.5*	70					
Modelo 5	13538.38*	4647			0.031	1.00	0.052	0.82	0.90
(5/3)			26.4 ns	22					
(5/1)			89.3 ns	67					

Tabla 6.5. Indices de bondad de ajuste de los modelos evaluados.

* = $p < 0.01$; ns = no significativo.

Modelo 1: Equivalencia estructural.

Modelo 2. Invarianza total de las saturaciones factoriales.

Modelo 3: Invarianza parcial de las saturaciones factoriales.

Modelo 4: Invarianza parcial de las saturaciones factoriales e invarianza total de los interceptos.

Modelo 5: Invarianza parcial de las saturaciones factoriales e invarianza parcial de los interceptos.

(2/1) = comparación de la diferencia de χ^2 del Modelo 2 respecto al Modelo 1.

(3/1) = comparación de la diferencia de χ^2 del Modelo 3 respecto al Modelo 1.

(4/3) = comparación de la diferencia de χ^2 del Modelo 4 respecto al Modelo 3.

(5/3) = comparación de la diferencia de χ^2 del Modelo 5 respecto al Modelo 3.

(5/1) = comparación de la diferencia de χ^2 del Modelo 5 respecto al Modelo 1.

En los siguientes apartados definimos y describimos cada uno de estos modelos, además de comentar los resultados obtenidos para cada uno de ellos en el análisis de bondad de ajuste.

2.1. Modelo 1: Equivalencia estructural

Para poner a prueba la equivalencia estructural de las dos versiones del instrumento, se llevó a cabo un análisis multimuestra con el programa LISREL 8 (Jöreskog y Sörbom, 1993). La matriz de correlaciones, y los vectores de desviaciones típicas y de medias de las variables observables para cada muestra se calcularon con el programa SPSS (1993) en su versión para Windows.

El modelo puesto a prueba asume que hay un total de 11 factores latentes correlacionados que subyacen a las 70 variables observables y que dan cuenta de las covarianzas observadas entre ellas. El análisis se realiza teniendo en cuenta únicamente las covarianzas entre las variables observables, es decir, se asume que tanto las medias de las variables observables como las medias de las variables latentes presentan un valor igual a cero (τ_X=0 y κ=0). Por lo tanto, las matrices que definen el modelo son las siguientes:

* Lambda-x (λ_X): es una matriz de orden (70 x 11) de saturaciones factoriales. Esta matriz fue definida para que cada variable observable saturara únicamente en un factor latente (aquel que según el modelo teórico debía medir), de este modo, sus saturaciones en las otras variables latentes presentaban un valor igual a 0. El procedimiento para fijar la escala de medida de cada una de las variables latentes

consistió en seleccionar de forma arbitraria una de las variables observables que saturaba en el factor, y fijar su saturación factorial a 1.

* Phi (Φ): es una matriz de orden (11 x 11) que representa las relaciones entre las variables latentes. Fue definida como una matriz simétrica, no diagonal y libre. Es decir, se estimaron los valores de las 11 varianzas de las variables latentes, y de las 55 covarianzas entre dichas variables latentes.

* Theta-Delta (Θ_δ): es una matriz de orden (70 x 70) que representa las relaciones entre los errores o unicidades. Esta matriz fue definida como una matriz diagonal y libre, es decir, se asume que los términos de error no están correlacionados, y por lo tanto todos los elementos por encima y por debajo de la diagonal presentan un valor igual a 0. Unicamente se estiman los valores de la diagonal, que representan las varianzas de cada término de error.

La figura 6.1. representa la estructura que se pone a prueba en este modelo en las dos muestras analizadas. Como ya hemos comentado anteriormente, las restricciones de equivalencia entre las dos muestras se refieren únicamente a la estructura pero no a los valores de los parámetros de las matrices definidas más arriba. Es decir, se pone a prueba si este patrón de relaciones entre variables observables, factores latentes y unicidades se ajusta a los datos en las dos muestras analizadas.

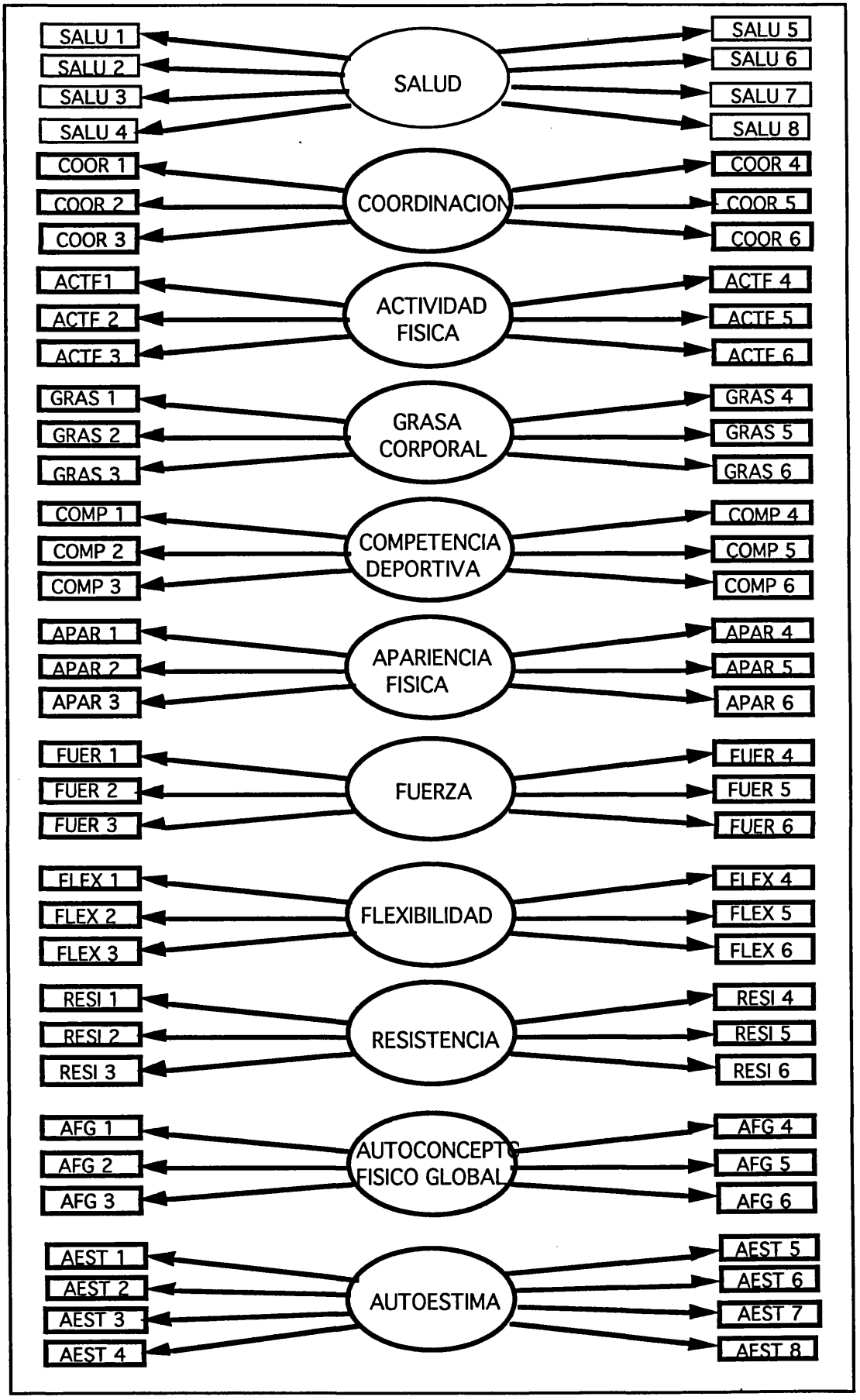


Fig. 6.1. Estructura factorial del Modelo 1.

Aunque el valor de χ^2 para este modelo (ver tabla 6.5.) es estadísticamente significativo (χ^2 (4580) = 13449.05, $p < 0.01$), los otros índices sugieren que el ajuste del modelo es razonable (RMSEA=0.031; RMSRS=0.053; GFI=0.82; y CFI=0.90). Hay que tener en cuenta que el estadístico χ^2 es sensible al tamaño de la muestra, por lo que incluso diferencias triviales entre la matriz de covarianzas reproducida $(\hat{\Sigma})$ y la matriz de covarianzas observada (S), pueden resultar estadísticamente significativas si el tamaño de la muestra es grande (Bentler y Bonett, 1980; Bollen, 1989), como es nuestro caso. Por lo tanto, el Modelo 1 fue tomado como base para evaluar el ajuste de los modelos más restrictivos que se pusieron a prueba posteriormente, y que pasamos a describir en los siguientes apartados.

2.2. Modelo 2: Invarianza total de las saturaciones factoriales

En este modelo, todos los parámetros de la matriz lambda-x son constreñidos a tomar el mismo valor en las dos muestras, es decir, se evalúa la invarianza de todas las saturaciones factoriales a través de los dos grupos culturales. Como este modelo se encuentra anidado en el Modelo 1, su ajuste puede ser evaluado en base a la diferencia de los valores del estadístico χ^2 de los dos modelos.

Como puede observarse en la tabla 6.5., aunque los índices de bondad de ajuste RMSEA (=0.032), RMSRS (=0.054), GFI (=0.81), y CFI (=0.89) son satisfactorios, la diferencia entre la χ^2 de este modelo y la del Modelo 1 es estadísticamente significativa (χ^2 (59) = 675.2, $p < 0.01$). Por lo tanto, no se confirma la hipótesis de la invarianza total de las saturaciones factoriales. Al no confirmarse esta hipótesis, se puede poner a prueba la invarianza parcial de las saturaciones factoriales. Este

modelo implica que algunas de las saturaciones factoriales de la matriz lambda-x son invariantes entre las dos muestras, pero no lo son todas. Por lo tanto, se debe identificar aquellos items cuya saturación factorial no es invariante y liberar sus parámetros lambda-x para mejorar el ajuste del modelo. Esto es lo que se hizo en el Modelo 3.

2.3. Modelo 3: Invarianza parcial de las saturaciones factoriales

Cuando un modelo no presenta un ajuste adecuado a los datos, los índices de modificación (IM, en adelante) ofrecen una información valiosa para decidir qué cambios se deben introducir en el modelo con el fin de mejorar su ajuste. El programa LISREL 8 proporciona un índice de modificación para cada parámetro fijado o constreñido en el modelo. El IM representa una estimación sobre cuánto disminuirá el valor de χ^2 del modelo si se libera ese parámetros en particular, y su valor está asociado con 1 grado de libertad.

Tras examinar los valores de los IM de los parámetros de la matriz lambda-x del Modelo 2, se detectaron dos items cuyos índices de modificación presentaban en las dos muestras, valores muy superiores a los de los otros items. Estos eran el ítem 1 de la escala "Flexibilidad" (IM=250.07), y el ítem 1 de la escala "Actividad física" (IM = 160.30). Para analizar la invarianza parcial se puso a prueba otro modelo con las mismas especificaciones que el Modelo 2, excepto que se liberaban las saturaciones factoriales de estos dos items, de tal forma que eran estimados libremente para cada grupo. Los valores del ajuste de este modelo presentaban una mejora respecto al Modelo 2, pero seguían

ofreciendo un valor estadísticamente significativo para la diferencia de las χ^2 respecto al Modelo 1.

Se fueron probando sucesivamente diferentes modelos anidados en los que se liberaban las saturaciones factoriales de aquellos items que presentaban mayores IM. Tras liberar un total de 14 parámetros de saturaciones factoriales se llegó a un modelo (Modelo 3), en el que la diferencia en los valores de χ^2 respecto al Modelo 1 no resultaba estadísticamente significativa (χ^2 (45) = 62.9, ns). Además, tal como puede verse en la tabla 6.5., los otros índices de ajuste (RMSEA=0.031; RMSRS=0.052; GFI=0.82; y CFI=0.90) presentan una ligera mejora respecto a los obtenidos para el Modelo 2, y ofrecen valores comparables a los obtenidos para el Modelo 1.

El Modelo 3 confirma la hipótesis de la invarianza parcial de las saturaciones factoriales, indicando que hay 56 items cuyas saturaciones son invariantes, lo que representa un 80% del total de items que componen el cuestionario. Los resultados relativos a los items que presentan FDI se recoge en el punto 3 "Análisis del funcionamiento diferencial de los items" de este capítulo.

El Modelo 3 ofrece información sobre la invarianza del parámetro de discriminación de los items. Para obtener información relativa a los interceptos se formuló otro modelo (Modelo 4) que definimos a continuación.

2.4. Modelo 4: Invarianza parcial de las saturaciones factoriales e invarianza total de los interceptos

Partiendo del Modelo 3 de invarianza parcial de las saturaciones factoriales, se puso a prueba otro modelo (Modelo 4) en el que a diferencia de los anteriores, se incorporan los valores de las medias de las variables observables, es decir, $\tau x \neq 0$. De este modo, el Modelo 4 queda definido por las 3 matrices descritas para los modelos anteriores (lambda-x, phi y theta-delta), y por dos vectores adicionales:

* Tau-x (τ_x): es un vector de orden (70 x 1) que contiene los interceptos de las 70 variables observables.

* Kappa (κ): es un vector de orden (11 x 1) que contiene la estimación de las medias de las variables latentes.

En el Modelo 4 la matriz lambda-x se define del mismo modo que en el Modelo 3, y por otra parte, todos los parámetros de la matriz tau-x son constreñidos a tomar el mismo valor en las dos muestras, es decir, se establece la invarianza parcial de las saturaciones factoriales, y la invarianza total de todos los interceptos a través de los dos grupos culturales. Como este modelo se encuentra anidado en el Modelo 3, su ajuste puede ser evaluado en base a la diferencia de los valores del estadístico χ^2 de los dos modelos.

En la tabla 6.5. se observa que aunque los índices de bondad de ajuste RMSEA (=0.033), RMSRS (=0.051), GFI (=0.81), y CFI (=0.88) son satisfactorios, la diferencia entre la χ^2 de este modelo y la del Modelo 3 es estadísticamente significativa (χ^2 (70) = 1399.5, $p < 0.01$). Por lo tanto,

no se confirma la hipótesis de la invarianza total de los interceptos. Igual que hemos visto anteriormente respecto a las saturaciones factoriales, al no confirmarse la hipótesis de la invarianza total de los interceptos se puede poner a prueba la hipótesis de la invarianza parcial. Este modelo implica que algunos de los interceptos de la matriz tau-x son invariantes entre las dos muestras, pero no lo son todos. En esta ocasión se deben identificar aquellos items cuyo intercepto no es invariante y liberar sus parámetros interceptos para mejorar el ajuste del modelo. Estas modificaciones fueron introducidas en el Modelo 5.

2.5. Modelo 5: Invarianza parcial de las saturaciones factoriales y de los interceptos

En base a los valores de los IM de los parámetros de la matriz tau-x se pusieron a prueba diferentes modelos con las mismas especificaciones que el Modelo 4, excepto que se liberaban los interceptos de aquellos items que presentaban mayores IM. Tras sucesivas modificaciones en las que se liberaban cierto número de interceptos, se estimaba el ajuste del modelo resultante, y se comparaba con el ajuste del Modelo 3, se llegó a un modelo (Modelo 5) en el que la diferencia en los valores de χ^2 respecto al Modelo 3 no resultaba estadísticamente significativa (χ^2 (22) = 26.4, ns). Para ello fue necesario liberar un total de 48 interceptos. Tal como puede verse en la tabla 6.5., los otros índices de ajuste (RMSEA=0.031; RMSRS=0.052; GFI=0.82; y CFI=0.90) presentan una ligera mejora respecto a los obtenidos para el Modelo 4, y ofrecen valores comparables a los obtenidos para el Modelo 3. Además, la diferencia en los valores de χ^2 del Modelo 5 respecto al modelo base (Modelo 1), tampoco resulta estadísticamente significativa (χ^2 (67) = 89.3, ns). Por lo tanto, el Modelo 5 indica que hay 22 items

cuyos interceptos son invariantes, lo que representa un 31,4% del total de items que componen el cuestionario.

La tabla 6.6. ofrece las saturaciones factoriales estimadas para el Modelo 5 con el programa LISREL 8. Todas las saturaciones factoriales son distintas de cero de forma estadísticamente significativa ($p < 0.05$). Para los 14 items cuya saturación factorial no es invariante en las dos versiones del cuestionario, se ofrece el valor obtenido con la versión en inglés en la muestra australiana (Aus.), y entre paréntesis el valor obtenido con la versión en castellano en la muestra española (Esp.). El símbolo (#) representa saturaciones factoriales a las que se ha asignado un valor igual a 1, con objeto de fijar la escala de medida de cada una de las variables latentes.

Los valores de los interceptos estimados para el Modelo 5 aparecen a continuación en la tabla 6.7. Todos los valores que aparecen en la tabla son distintos de cero de forma estadísticamente significativa ($p < 0.05$). Para los 48 items cuyo intercepto no es invariante, se ofrece el valor obtenido en cada una de las dos muestras.

Finalmente, en el apartado 3 presentamos los resultados relativos a los items que presentan funcionamiento diferencial en las dos versiones del PSDQ, y se analiza el tipo de FDI que presentan.

Tabla 6.6. Saturaciones factoriales estimadas para el Modelo 5. (Continúa).

		Subescalas										
Items		SALU	COOR	ACTF	GRAS	COMP	APAR	FUER	FLEX	RESI	AFG	AEST
Salu1		0.50										
Salu2		1.00#										
Salu3		0.71										
Salu4		1.02										
Salu5		1.02										
Salu6		0.87										
Salu7	Aus.	0.90										
	Esp.	(1.06)										
Salu8		0.76										
Coor1			0.75									
Coor2			0.69									
Coor3			1.00#									
Coor4	Aus.		1.01									
	Esp.		(0.88)									
Coor5	Aus.		0.91									
	Esp.		(1.11)									
Coor6			0.97									
Actf1	Aus.			0.80								
	Esp.			(0.38)								
Actf2	Aus.			0.72								
	Esp.			(0.95)								
Actf3	Aus.			0.83								
	Esp.			(1.10)								
Actf4				0.94								
Actf5				1.00#								
Actf6				1.06								
Gras1					1.15							
Gras2	Aus.				1.10							
	Esp.				(1.00)							
Gras3					1.22							
Gras4					1.20							
Gras5					1.13							
Gras6					1.00#							
Comp1						0.95						
Comp2						1.00						
Comp3	Aus.					0.88						
	Esp.					(0.78)						
Comp4						1.03						
Comp5						0.90						
Comp6						1.00#						
Apar1							1.00#					
Apar2	Aus.						1.01					
	Esp.						(0.86)					
Apar3							1.04					
Apar4							1.02					
Apar5							1.15					
Apar6	Aus.						0.76					
	Esp.						(0.64)					

Tabla 6.6. Saturaciones factoriales estimadas para el Modelo 5. (continuación).

Items		SALU	COOR	ACTF	GRAS	COMP	APAR	FUER	FLEX	RESI	AFG	AEST
Fuer1								1.05				
Fuer2								1.10				
Fuer3	Aus.							1.15				
	Esp.							(1.02)				
Fuer4								0.66				
Fuer5								1.13				
Fuer6								1.00#				
Flex1	Aus.								1.05			
	Esp.								(0.38)			
Flex2									1.19			
Flex3									0.60			
Flex4									0.84			
Flex5									1.00#			
Flex6									1.15			
Resi1										1.00#		
Resi2	Aus.									0.82		
	Esp.									0.97		
Resi3										1.03		
Resi4										1.06		
Resi5										0.90		
Resi6										0.94		
Afg1											0.89	
Afg2											0.98	
Afg3											1.00#	
Afg4											1.08	
Afg5	Aus.										1.02	
	Esp.										(0.89)	
Afg6											1.08	
Aest1												0.67
Aest2												1.00
Aest3												1.01
Aest4												1.05
Aest5												0.84
Aest6												1.18
Aest7												1.00#
Aest8												1.02

Todas las saturaciones son distintas de cero, $p < 0.05$.

Aus. = saturación en la muestra australiana de los items cuya saturación no es invariante

Esp. = saturación en la muestra española de los items cuya saturación no es invariante. Valor entre paréntesis ( ).

\# = saturaciones factoriales con valor asignado igual a 1 para fijar la escala de medida de cada una de las variables latentes.

Tabla 6.7. Interceptos estimados para el Modelo 5.

Item	Muestra australiana	Muestra española
Salu1	4.36	4.22
Salu2	4.36	4.79
Salu3	5.28	
Salu4	3.98	4.16
Salu5	5.10	
Salu6	4.90	4.73
Salu7	5.17	4.97
Salu8	4.41	
Coor1	4.43	4.59
Coor2	4.66	4.53
Coor3	4.31	
Coor4	4.34	
Coor5	4.25	
Coor6	4.21	
Actf1	4.46	4.13
Actf2	4.44	4.18
Actf3	3.93	
Actf4	4.45	
Actf5	4.42	4.21
Actf6	3.92	
Gras1	4.74	4.60
Gras2	4.58	4.31
Gras3	4.49	
Gras4	4.79	4.40
Gras5	4.60	
Gras6	4.88	4.76
Comp1	3.79	
Comp2	4.39	
Comp3	4.39	4.47
Comp4	4.38	
Comp5	3.69	3.56
Comp6	4.53	4.43
Apar1	3.53	
Apar2	3.60	4.02
Apar3	3.24	3.05

Item	Muestra australiana	Muestra española
Apar4	4.46	4.08
Apar5	3.66	
Apar6	4.18	
Fuer1	4.21	3.99
Fuer2	4.28	3.86
Fuer3	3.59	3.27
Fuer4	4.96	4.51
Fuer5	3.94	3.75
Fuer6	4.12	3.69
Flex1	4.14	3.29
Flex2	3.75	3.57
Flex3	4.76	4.39
Flex4	4.51	4.22
Flex5	4.42	4.04
Flex6	3.76	3.57
Resi1	3.79	
Resi2	3.81	3.96
Resi3	2.99	3.11
Resi4	3.48	3.64
Resi5	4.01	3.95
Resi6	3.96	4.12
Afg1	4.52	4.71
Afg2	4.60	4.70
Afg3	4.23	4.42
Afg4	4.49	
Afg5	4.60	4.73
Afg6	4.55	4.40
Aest1	4.45	4.04
Aest2	4.57	4.14
Aest3	4.73	
Aest4	5.11	
Aest5	4.64	4.11
Aest6	4.57	4.37
Aest7	5.20	5.02
Aest8	4.76	4.60

2. ANALISIS DEL FUNCIONAMIENTO DIFERENCIAL DE LOS ITEMS

Los diferentes modelos puestos a prueba y que han sido expuestos en el apartado anterior, permiten analizar la invarianza (en las dos versiones del instrumento) de los parámetros de discriminación y de los "parámetros de dificultad" de los items que componen el PSDQ, y por lo tanto permiten determinar qué items presentan FDI y qué tipo de FDI presentan. Hemos entrecomillado el término parámetro de dificultad porque somos conscientes de que esta denominación común que desde el enfoque de la TRI se da al parámetro b de los items, no resulta del todo adecuada en el tipo de cuestionario que se analiza en este trabajo. El término parámetro de dificultad surge en el contexto de los tests de aptitudes y rendimiento, donde las respuestas son definidadas como aciertos o errores, y el índice de dificultad representa el nivel de aptitud en el punto de inflexión de la CCI. En este contexto, la dificultad del ítem describe dónde está situado el ítem en la escala de aptitud, es decir, qué cantidad de aptitud requiere el ítem para ser resuelto con éxito (Martínez Arias, 1995). Esta interpretación requiere ser adaptada al contexto de la medida de otros constructos psicológicos, como puede ser el autoconcepto físico, ya que las respuestas no indican aciertos ni errores, y el parámetro de dificultad refleja diferentes niveles de respuesta evocados por la situación particular a la que hace referencia el ítem (Ferrando, 1996b). Se han propuesto diferentes nombres para hacer referencia al parámetro b dentro de este contexto, como por ejemplo el de "parámetro de preferencia" (preference parameter), el de "parámetro de atraccción" (attractiveness parameter), o el de "parámetro de evocación" (evocativeness parameter) (Lanning, 1991). Con esta aclaración queremos resaltar que dada la naturaleza del instrumento de medida analizado en este trabjao, al estudiar y describir

los items que lo componen, quizá sería más adecuado utilizar el término de parámetro de evocación en lugar del de parámetro de dificultad. Sin embargo, ya que lo hemos utilizado hasta ahora, vamos a seguir utilizando este último término, asumiendo implícitamente los matices que implican hablar del parámetro b en este contexto de medida de constructos psicológicos.

Tal como se ha justificado en el capítulo 3, la saturación factorial puede ser interpretada como el parámetro de discriminación, y el intercepto como el parámetro de dificultad del ítem. De este modo, la presencia o ausencia de invarianza de estos parámetros nos permite clasificar los 70 items del PSDQ en tres categorías:

1) Items que presentan FDI no uniforme: la saturación factorial no es invariante, independientemente de que el intercepto sea invariante o no lo sea. Se han detectado un total de 14 items que presentan FDI no uniforme, lo que representa un 20% del total de los items que componen el cuestionario. En la tabla 6.8. aparecen los items que presentan FDI no uniforme agrupados por subescalas. El número que figura delante de cada ítem corresponde a su numeración en el cuestionario. La clave que figura al final entre paréntesis indica su numeración dentro de la escala a la que pertenece; dicha numeración corresponde al orden en el que aparecen los items en la tabla 5.1.

Tabla 6.8. Items que presentan FDI no uniforme.

SALUD

67. Me pongo enfermo/a y tengo que ir al médico con más frecuencia que la mayoría de los chicos/as de mi edad. (Salu7)

COORDINACION

35. En la mayoría de las actividades físicas, puedo realizar los movimientos con armonía. (Coor4)
46. Realizo con facilidad movimientos que requieren coordinación. (Coor5)

ACTIVIDAD FISICA

3. Varias veces a la semana realizo ejercicios o deportes lo suficientemente duros como para hacerme respirar fuerte. (Actf1)
14. Suelo hacer ejercicio o actividades que me hacen respirar fuerte. (Actf2)
25. Tres o cuatro veces a la semana y al menos durante media hora, hago ejercicio o actividades que me hacen respirar fuerte. (Actf3)

GRASA CORPORAL

15. Mi cintura es demasiado ancha. (Gras2)

COMPETENCIA DEPORTIVA

27. La mayoría de deportes me resultan fáciles. (Comp3)

APARIENCIA FISICA

18. Tengo una cara agradable. (Apar2)
62. Nadie piensa que soy guapo/a. (Apar6)

FUERZA

30. Soy más fuerte que la mayoría de los chicos/as de mi edad. (Fuer3)

FLEXIBILIDAD

9. Soy bastante bueno/a doblándome y retorciendo mi cuerpo. (Flex1)

RESISTENCIA

21. Obtendría buenos resultados en una prueba de resistencia física. (Resi2)

AUTOCONCEPTO FISICO GLOBAL

50. Me siento satisfecho/a con quien soy y con lo que puedo hacer físicamente. (Glph5)

2) Items que presentan FDI uniforme: la saturación factorial es invariante, pero el intercepto no es invariante. Se han detectado un total de 38 items que presentan FDI uniforme, lo que representa un 54.3% del total de los items que componen el cuestionario. En la tabla 6.9. aparecen los items que presentan FDI uniforme agrupados por subescalas.

Tabla 6.9. Items que presentan FDI uniforme. (Continúa)

SALUD

1. Cuando estoy enfermo/a, me encuentro tan mal que no puedo ni levantarme de la cama. (Salu1)
12. Normalmente cojo todas las enfermedades (gripe, virus, resfriados, etc.) que hay por ahí. (Salu2)
34. Casi nunca me pongo enfermo/a. (Salu4)
56. Cuando me pongo enfermo/a me cuesta mucho tiempo recuperarme. (Salu6)

COORDINACION

2.Me siento seguro/a realizando movimientos que requieren coordinación. (Coor1)
13. Me resulta fácil controlar los movimientos de mi cuerpo. (Coor2)

ACTIVIDAD FISICA

47. Practico muchos deportes, baile, gimnasia u otras actividades físicas. (Actf5)

GRASA CORPORAL

4. Estoy demasiado gordo/a. (Gras1)
37. Peso demasiado. (Gras4)
59. La gente piensa que estoy gordo/a. (Gras6)

COMPETENCIA DEPORTIVA

49. Se me dan mejor los deportes que a la mayoría de mis amigos/as. (Comp5)
60. Juego bien en los deportes. (Comp6)

APARIENCIA FISICA

29. Soy más guapo/a que la mayoría de mis amigos/as. (Apar3)
40. Soy feo/a. (Apar4)

FUERZA

8. Soy una persona físicamente fuerte. (Fuer1)
19. Tengo mucha fuerza física. (Fuer2)
41. Soy débil y casi no tengo músculo. (Fuer4)
52. Obtendría buenos resultados en una prueba de fuerza. (Fuer5)
63. Se me da bien levantar objetos pesados. (Fuer6)

FLEXIBILIDAD

20. Mi cuerpo es flexible. (Flex2)
31. Mi cuerpo es rígido y nada flexible. (Flex3)
42. Puedo doblar y mover bien las diversas partes de mi cuerpo en la mayoría de las direcciones. (Flex4)
53. Creo que tengo bastante flexibilidad para la práctica de la mayoría de los deportes. (Flex5)
64. Creo que obtendría buenos resultados en una prueba de flexibilidad. (Flex6)

RESISTENCIA

32. Podría correr durante 5 kilómetros sin parar. (Resi3)
43. Creo que podría correr una distancia larga sin cansarme. (Resi4)
54. Puedo mantenerme físicamente activo/a durante un periodo largo de tiempo sin cansarme. (Resi5)
65. Se me dan bien las actividades de resistencia física, como las carreras de larga distancia, el aerobic, el ciclismo o la natación. (Resi6)

AUTOCONCEPTO FISICO GLOBAL

6. Físicamente, estoy satisfecho/a con el tipo de persona que soy. (Afg1)
17. Físicamente, me siento contento/a conmigo mismo/a. (Afg2)
28. Me siento satisfecho/a con mi apariencia física y con lo que puedo hacer físicamente. (Afg3)
61. Estoy satisfecho/a con cómo soy físicamente. (Afg6)

AUTOESTIMA

11. En general, la mayoría de las cosas que hago me salen bien. (Aest1)
22. No tengo mucho de lo que sentirme orgulloso/a. (Aest2)
55. Hago bien la mayoría de las cosas que hago. (Aest5)
66. En general, tengo mucho de lo que sentirme orgulloso/a. (Aest6)
68. En general, soy un fracaso. (Aest7)
70. Nada de lo que hago parece salir bien. (Aest8)

3) Items que no presentan FDI: tanto la saturación factorial como el intercepto son invariantes en las dos versiones del instrumento. Se han detectado un total de 18 items que no presentan FDI, lo que representa un 25.7% del total de los items que componen el cuestionario. En la tabla 6.10. aparecen los items que no presentan FDI.

Tabla 6.10. Items que no presentan FDI.

SALUD

23. Estoy enfermo/a tan a menudo que no puedo hacer todas las cosas que quisiera. (Salu3)
45. Me pongo enfermo/a con mucha frecuencia. (Salu5)
69. Normalmente me mantengo sano/a, incluso cuando mis amigos/as se ponen enfermos. (Salu8)

COORDINACION

24. Soy bueno/a realizando movimientos que requieren coordinación. (Coor3)
57. Me muevo con gracia y coordinación cuando practico deportes y actividades. (Coor6)

ACTIVIDAD FISICA

36. Hago actividades físicas (como correr, bailar, ir en bici, aerobic, gimnasia o nadar) por lo menos tres veces a la semana. (Actf4)
58. Practico deportes, ejercicio, baile u otras actividades físicas casi todos los días. (Actf6)

GRASA CORPORAL

26. Tengo demasiada grasa en mi cuerpo. (Gras3)
48. Mi barriga es demasiado grande. (Gras5)

COMPETENCIA DEPORTIVA

5. La gente piensa que soy bueno/a en los deportes. (Comp1)
16. Se me dan bien la mayoría de deportes. (Comp2)
38. Tengo buenas habilidades deportivas. (Comp4)

APARIENCIA FISICA

7. Teniendo en cuenta mi edad, soy atractivo/a. (Apar1)
51. Soy guapo/a. (Apar5)

RESISTENCIA

10. Puedo correr largas distancias sin parar. (Resi1)

AUTOCONCEPTO FISICO GLOBAL

39. Físicamente, me siento satisfecho/a conmigo mismo/a. (Afg4)

AUTOESTIMA

33. Siento que mi vida no es demasiado útil. (Aest3)
44. En general, no valgo para nada. (Aest4)

En la tabla 6.11. presentamos un resumen del funcionamiento de los items dentro de cada escala.

Tabla 6.11. Tabla resumen del FDI presentado por los items del PSDQ. (Continúa).

Items	Tipo de FDI	FDI Unif. a favor de:
Salud		
Salu1	FDI Unif.	(+ Aust.)
Salu2	FDI Unif.	(+ Esp.)
Salu3	--	
Salu4	FDI Unif.	(+ Esp.)
Salu5	--	
Salu6	FDI Unif.	(+ Aust.)
Salu7	FDI No Uni.	
Salu8	--	
Coordinación		
Coor1	FDI Unif.	(+ Esp.)
Coor2	FDI Unif.	(+ Aust.)
Coor3	--	
Coor4	FDI No Uni.	
Coor5	FDI No Uni.	
Coor6	--	
Actividad Física		
Actf1	FDI No Uni.	
Actf2	FDI No Uni.	
Actf3	FDI No Uni.	
Actf4	--	
Actf5	FDI Unif.	(+ Aust.)
Actf6	--	
Grasa Corporal		
Gras1	FDI Unif.	(+ Aust.)
Gras2	FDI No Uni.	
Gras3	--	
Gras4	FDI Unif.	(+ Aust.)
Gras5	--	
Gras6	FDI Unif.	(+ Aust.)

Items	Tipo de FDI	FDI Unif. a favor de:
Comp.Dep.		
Comp1	--	
Comp2	--	
Comp3	FDI No Uni.	
Comp4	--	
Comp5	FDI Unif.	(+ Aust.)
Comp6	FDI Unif.	(+ Aust.)
Apariencia		
Apar1	--	
Apar2	FDI No Uni.	
Apar3	FDI Unif.	(+ Aust.)
Apar4	FDI Unif.	(+ Aust.)
Apar5	--	
Apar6	FDI No Uni.	
Fuerza		
Fuer1	FDI Unif.	(+ Aust.)
Fuer2	FDI Unif.	(+ Aust.)
Fuer3	FDI No Uni.	
Fuer4	FDI Unif.	(+ Aust.)
Fuer5	FDI Unif.	(+ Aust.)
Fuer6	FDI Unif.	(+ Aust.)
Flexibilidad		
Flex1	FDI No Uni.	
Flex2	FDI Unif.	(+ Aust.)
Flex3	FDI Unif.	(+ Aust.)
Flex4	FDI Unif.	(+ Aust.)
Flex5	FDI Unif.	(+ Aust.)
Flex6	FDI Unif.	(+ Aust.)

Tabla 6.11. Tabla resumen del FDI presentado por los items del PSDQ. (continuación).

Items	Tipo de FDI	FDI Unif. a favor de:
Resistencia		
Resi1	--	
Resi2	FDI No Uni.	
Resi3	FDI Unif.	(+ Esp.)
Resi4	FDI Unif.	(+ Esp.)
Resi5	FDI Unif.	(+ Aust.)
Resi6	FDI Unif.	(+ Esp.)
Autoestima		
Aest1	FDI Unif.	(+ Aust.)
Aest2	FDI Unif.	(+ Aust.)
Aest3	--	
Aest4	--	
Aest5	FDI Unif.	(+ Aust.)
Aest6	FDI Unif.	(+ Aust.)
Aest7	FDI Unif.	(+ Aust.)
Aest8	FDI Unif.	(+ Aust.)

Items	Tipo de FDI	FDI Unif. a favor de:
Autocon. Físico Global		
Afg1	FDI Unif.	(+ Esp.)
Afg2	FDI Unif.	(+ Esp.)
Afg3	FDI Unif.	(+ Esp.)
Afg4	--	
Afg5	FDI No Uni.	
Afg6	FDI Unif.	(+ Aust.)

Clave: FDI No Uni. = items que presentan funcionamiento diferencial no uniforme.
FDI Unif. = items que presentan funcionamiento diferencial uniforme.
(Aust.) = FDI uniforme a favor de los sujetos de la muestra australiana.
(Esp.) = FDI uniforme a favor de los sujetos de la muestra española.
-- = items que no presentan funcionamiento diferencial.

En esta tabla aparece indicado los items que presentan FDI y el tipo de FDI que presentan. Para aquellos items que presentan FDI uniforme se indica si éste es a favor de la muestra australiana (Aust.), o a favor de la muestra española (Esp.).

Tal como puede verse en la tabla 6.11., hay dos subescalas (Fuerza y Flexibilidad) en las que todos los items presentan FDI, uniforme o no uniforme. En las otras subescalas, el número de items que no presenta

FDI oscila entre 1 y 3. En la tabla 6.12. presentamos el número y el porcentaje de items que no presentan FDI para cada subescala.

Subescalas:		
Competencia Deport.	3 items sin FDI	50.0 % del nº total de items
Salud	3 items " "	37.5 % " " "
Coordinación	2 items " "	33.3 % " " "
Actividad Física	2 items " "	33.3 % " " "
Grasa Corporal	2 items " "	33.3 % " " "
Apariencia Física	2 items " "	33.3 % " " "
Apariencia Física	2 items " "	33.3 % " " "
Autoestima	2 items " "	25.0 % " " "
Resistencia	1 ítem " "	16.6 % " " "
Resistencia	1 ítem " "	16.6 % " " "
Autocon.Fisic.Global	1 ítem " "	16.6 % " " "
Fuerza	0 items " "	0.0 % " " "
Flexibilidad	0 items " "	0.0 % " " "

Tabla 6.12. Número de items que no presentan FDI y porcentaje que representa del número total de items de cada subescala.

También podemos observar en la tabla 6.11. que de los items que presentan FDI uniforme, hay una mayor tendencia a que éste sea a favor de la muestra australiana. Esto es especialmente evidente en las subescalas Autoestima, Fuerza, Flexibilidad, Grasa Corporal, Apariencia Física, Competencia Deportiva, y Actividad Física, en las que todos los items que muestran FDI uniforme lo hacen a favor de la muestra australiana. En las otras subescalas hay items con FDI uniforme a favor de la muestra australiana, e items con FDI uniforme a favor de la muestra española. En la tabla 6.13. detallamos la proporción de items

que presentan FDI uniforme a favor de un grupo u otro para cada subescala. La notación utilizada representa el número de items totales que muestran FDI uniforme en cada subescala, el número de estos que indican una ventaja a favor de la muestra australiana o española, y el porcentaje que éstos representan del total. Por ejemplo, en la subescala resistencia hay un total de 4 items que presentan FDI uniforme, de los cuales 1 lo hace a favor de la muestra australiana (1/4) lo que representa el 25% del número total de items con FDI uniforme de esta subescala, y 3 lo hacen a favor de la muestra española (3/4) lo que representa el 75% de los items con FDI uniforme de esta subescala.

Subescala	Items con FDI uniforme a favor de la muestra australiana		Items con FDI uniforme a favor de la muestra española	
Autoestima	(6/6)	100 %		
Fuerza	(5/5)	100 %		
Flexibilidad	(5/5)	100 %		
Grasa Corporal	(3/3)	100 %		
Aparien.Física	(2/2)	100 %		
Comp.Deporti.	(2/2)	100 %		
Actívid.Física	(1/1)	100 %		
Salud	(2/4)	50 %	(2/4)	50 %
Coordinación	(1/2)	50 %	(1/2)	50 %
Resistencia	(1/4)	25 %	(3/4)	75 %
Autoc.Fis.Glob.	(1/4)	25 %	(3/4)	75 %

Tabla 6.13. Proporción de items que presentan FDI uniforme en cada escala a favor de la muestra australiana y a favor de la muestra española.

En el capítulo 7 pasamos a discutir y a reflexionar sobre los resultados obtenidos en este estudio. Comentaremos también sus

aportaciones, limitaciones, y las futuras investigaciones que se pueden plantear a partir de este trabajo.

CAPITULO 7

DISCUSION Y CONCLUSIONES

El análisis de los resultados presentados en el capítulo 6 pone de manifiesto que al menos en parte, se ha logrado el objetivo planteado en este estudio, y aquellos aspectos que no han podido ser alcanzados, abren nuevas vías de investigación y de análisis, así como interrogantes y llamadas de atención sobre el proceso de traducción de instrumentos de medida y sobre los procedimientos utilizados en los estudios de comparación transcultural entre grupos que no comparten el mismo idioma. Recordemos que el objetivo general de este trabajo era ofrecer una versión en castellano del PSDQ que fuera equivalente a la versión original en inglés del cuestionario, de tal forma que permitiera la comparación transcultural del autoconcepto físico entre muestras de sujetos que pertenecen a diferentes culturas y que no comparten el mismo idioma (inglés versus castellano). Este objetivo general se desglosaba en tres objetivos específicos.

El primero de los objetivos específicos era llevar a cabo una traducción y adaptación al castellano del PSDQ utilizando los métodos de juicio y los consejos y recomendaciones que recoge la literatura transcultural para garantizar la adecuada traducción del instrumento. Este es uno de los logros conseguidos en el presente trabajo. El riguroso proceso de traducción y adaptación seguido para elaborar la versión en castellano del PSDQ nos lleva a resaltar la calidad de esta nueva versión del cuestionario. Además, cabe señalar que las 11 subescalas que componen esta versión traducida, presentan valores satisfactorios de consistencia interna, siendo éstos similares a los presentados por las subescalas de la versión original. Podemos concluir por tanto, que el primero de los objetivos específicos planteados ha sido alcanzado satisfactoriamente. El siguiente paso, que se plantea en los dos objetivos específicos posteriores, es analizar si esta versión del instrumento es adecuada para ser utilizada en la investigación transcultural. Es decir, si ofrece medidas equivalentes del autoconcepto físico que garantizen la validez de los resultados en estudios de comparación entre grupos de sujetos que pertenecen a diferentes culturas y que no comparten el mismo idioma.

El segundo objetivo específico planteado en este trabajo era analizar la equivalencia estructural de la versión traducida al castellano del PSDQ respecto a la versión original en inglés, para comprobar si ambas miden el mismo constructo y con las mismas dimensiones. Los resultados de este estudio confirman la equivalencia estructural de las dos versiones. Un modelo de 11 factores que representan 9 componentes específicos del autoconcepto físico (salud, coordinación, actividad física, grasa corporal, competencia deportiva, apariencia física, fuerza, flexibilidad, y resistencia) y 2 componentes globales (autoconcepto físico

global, y autoestima), presenta un ajuste satisfactorio en las dos muestras analizadas. Por lo tanto, los resultados obtenidos respecto al segundo objetivo resultan plenamente satisfactorios, e indican una primera aproximación hacia el objetivo general planteado: ofrecer una versión en castellano del PSDQ equivalente a la versión original en inglés, que permita la comparación transcultural. De forma adicional, estos resultados ofrecen evidencia a favor de la validez de constructo del PSDQ.

Finalmente, el tercer objetivo específico planteado era analizar la invarianza factorial de las dos versiones del instrumento. Esto se hizo planteando diferentes hipótesis de invarianza respecto a las saturaciones factoriales y a los interceptos de los items que componen el cuestionario. En esta ocasión, los resultados obtenidos no son tan satisfactorios, ya que se han detectado un total de 52 items que presentan un funcionamiento diferencial en las dos versiones, lo que representa un 74.2% del total de los items que componen el cuestionario. Ante estos resultados, parecería necesario concluir que el tercer objetivo, y en consecuencia, el objetivo general planteado en este trabajo no se han visto confirmados: los resultados ofrecen evidencia en contra de la invarianza de alguno de los parámetros de un gran número de items que componen el cuestionario, lo que llevaría a cuestionar seriamente la equivalencia de las dos versiones del PSDQ, y por lo tanto, el que la versión desarrollada en castellano pudiera ser utilizada para la comparación transcultural. Sin embargo, antes de concluir lo anterior, es necesario introducir una serie de matizaciones que presentan una panorámica menos pesimista sobre los resultados obtenidos.

En primer lugar hay que señalar que el único requisito para poder obtener y comparar las medias de las variables latentes de diferentes grupos, es que se confirme la invarianza total o parcial de las saturaciones factoriales de los items (Byrne, Shavelson, y Muthèn, 1989). Aquellas subescalas en las que todos los items presentan invarianza de las saturaciones factoriales, ofrecen directamente una métrica común, lo que permite comparar las puntuaciones obtenidas en dichos factores latentes con las diferentes versiones del instrumento. Por otro lado, cuando las subescalas presentan invarianza parcial de las saturaciones factoriales, el conjunto de items deben ser equiparados a una métrica común antes de poder estimar valores comparables de las medias de las variables latentes en los diferentes grupos. En nuestro caso, el análisis del ajuste de diferentes modelos anidados ha confirmado la hipótesis de la invarianza parcial de las saturaciones factoriales. La confirmación de esta hipótesis implica que aquellos valores de lambda-x que son invariantes en los dos grupos comparados permiten definir la métrica de las variables latentes. La media en el factor puede ser estimada aplicando ponderaciones comunes a los items invariantes, y ponderaciones diferentes en función del grupo de pertenencia, a los items no invariantes. De este modo, las medias obtenidas para las variables latentes, están en una métrica común, y por lo tanto son comparables. Para asegurar que la comparación entre los grupos no es arbitraria, es necesario que la mayor parte de los items de cada variable latente presenten saturaciones factoriales invariantes (Reise, Widaman, y Pugh, 1993).

El problema empírico que se plantea es determinar cuántos items con saturaciones factoriales invariantes se requieren para justificar el empleo de métodos de equiparación a una métrica común, y asegurar

que la comparación entre los grupos no va a resultar arbitraria (Hulin y Mayer, 1986). Estos autores señalan la necesidad de realizar estudios de simulación en los que se manipule el número y la proporción total de items equivalentes y no equivalentes que componen una escala, y se analicen los efectos de esta manipulación sobre equivalencia métrica.

En base a los resultados de este estudio, parece bastante razonable concluir que al menos 8 de las 11 subescalas que componen la versión traducida del PSDQ (salud, grasa corporal, competencia deportiva, fuerza, flexibilidad, resistencia, autoconcepto físico global, y autoestima) presentan un número suficiente de items con invarianza de las saturaciones factoriales, para poder llevar a cabo la equiparación a una métrica común y permitir la comparación entre grupos de las medias de estos factores. El porcentaje de items de estas escalas que presentan invarianza de la saturación factorial oscila entre el 83.3%. y el 100%. La aceptabilidad de estos resultados se basa en la comparación con los obtenidos en otros estudios, como por ejemplo el de Reise y cols. (1993), en el que los autores concluyen sobre lo adecuado de establecer una métrica común para posibilitar comparaciones entre grupos, en una escala en la que el 80% de los items presentan invarianza de las saturaciones factoriales. Respecto a las otras tres escalas, que presentan un 66.7% (coordinación y apariencia física) y un 50% (actividad física) de items con invarianza factorial, aunque es posible establecer una métrica común en base a los items invariantes que contienen, el problema reside en determinar si esto sería correcto desde un punto de vista psicométrico. Por lo tanto, se puede concluir que se han obtenido un grupo de items con saturaciones invariantes, los cuales permiten estimar y comparar entre diferentes grupos culturales, al menos la mayor parte de las dimensiones del autoconcepto físico medidas por el PSDQ.

Otra de las aportaciones de este estudio ha sido la aplicación del AFC con estructura de medias latentes en el estudio de la equivalencia de diferentes versiones de un instrumento de medida. Ya hemos comentado anteriormente la reducida difusión del empleo de esta metodología; además, en los trabajos en los que ha sido utilizada, el interés se ha centrado en el establecimiento de la equivalencia entre grupos en función de variables como el sexo (por ejemplo: Everson, Millsap y Rodriguez, 1991; Ferrando, 1996), o el nivel académico (Byrne, Shavelson, y Muthén, 1989). Otras metodologías como los procedimientos de la TRI o el AFC de estructuras de covarianzas, han sido ampliamente utilizadas en el estudio de la equivalencia de diferentes versiones de un instrumento para su uso en estudios de comparación transcultural. Sin embargo, no conocemos ningún estudio en el que se haya utilizado el AFC con estructura de medias latentes con ese propósito. Por lo tanto, este trabajo ofrece evidencia de la utilidad del AFC con medias latentes para el estudio de la equivalencia de diferentes versiones instrumentos de medida.

El uso del AFC con estructura de medias latentes ha permitido identificar un total de 18 items cuyas saturaciones factoriales e interceptos son invariantes a través de los grupos. Esto significa que se ha conseguido un conjunto de items calibrados, cuyos parámetros son conocidos y pueden ser utilizados para puntuar a los sujetos. Si se utiliza este conjunto de items para medir a sujetos pertenecientes a diferentes grupos, se puede considerar que la escala de medida es la misma, y que las puntuaciones estimadas en el rasgo son comparables (Reise y cols. 1993).

El paso que permite dar el AFC con medias latentes respecto al AFC con estructuras de covarianza, es la consideración de los parámetros interceptos, y por tanto la posibilidad de analizar el FDI uniforme además del no uniforme que ya era detectable con el otro procedimiento. El uso de esta metodología en el presente trabajo, ha supuesto también la identificación de un mayor número de items que presentaban FDI. Este hecho merece dos comentarios. En primer lugar, pone de manifiesto la no detección de items que realmente presentan FDI (en este caso FDI uniforme) cuando se realizan análisis utilizando únicamente la estructura de covarianzas. Por otro lado, el elevado número de items detectados que presentan FDI uniforme, ha puesto en un primer momento en evidencia la equivalencia de la versión en castellano del cuestionario. Sin embargo, también aquí se hace necesario matizar, que estos resultados pueden estar en parte influidos por la relación entre la potencia del test χ^2 y el elevado tamaño de las muestras analizadas. Esto nos plantea la posibilidad de futuras ampliaciones de este trabajo que impliquen analizar el funcionamiento diferencial de los items con otros métodos de DFDI, como por ejemplo procedimientos basados en la TRI empleando el modelo de respuesta graduada de Samejima (1969), y comparar los resultados obtenidos con los diferente métodos utilizados.

Resumiendo, en este trabajo se ha analizado la equivalencia de la versión en castellano del PSDQ desde dos perspectivas diferentes y complementarias: una macroperspectiva, que representa el estudio de la equivalencia de la estructura factorial del cuestionario en sus dos versiones (equivalencia estructural); y una microperspectiva, que representa el estudio de la equivalencia de los items que componen las diferentes versiones del cuestionario (invarianza factorial). La

equivalencia ha sido confirmada a un nivel macro (las dos versiones presentan la misma estructura factorial), pero no ha sido confirmada a un nivel micro (algunos items no son equivalentes en las dos versiones del PSDQ). No obstante, la obtención de un grupo de items invariantes, señala la posibilidad de realizar comparaciones transculturales del autoconcepto físico en base a la versión en castellano del PSDQ elaborada.

Estos resultados nos llevan, por un lado a una reflexión metodológica, y por otro a plantear la necesidad de profundizar en las causas de los resultados encontrados. La reflexión metodológica sería también una llamada de atención sobre los procedimientos utilizados en la traducción y adaptación de cuestionarios, y en los estudios de comparación transcultural. En muchas ocasiones, no se tienen en cuenta ni los métodos de juicio ni los consejos y recomendaciones que recoge la literatura transcultural para garantizar la adecuada traducción de los instrumentos de medida. Además, se realizan comparaciones transculturales asumiendo la equivalencia de las diferentes versiones del cuestionario, pero sin llegar a realizar análisis que confirmen dicha equivalencia (Van de Vijver y Leung, 1996). Esto nos lleva a cuestionar seriamente la calidad de las traducciones obtenidas, y la validez de los resultados de estos estudios transculturales. Respecto a los procedimientos utilizados en la traducción de instrumentos de medida, este estudio es un claro ejemplo de que incluso llevando a cabo un proceso riguroso en la traducción, no siempre se consigue obtener versiones estrictamente equivalentes del cuestionario. Por otro lado, realizar estudios de comparación transcultural con instrumentos cuya equivalencia no ha sido analizada implica un riesgo respecto a la validez de los resultados encontrados. Por lo tanto, los resultados de este estudio

nos llevan, no solamente a plantear la necesidad de utilizar los métodos necesarios que garanticen la calidad de la traducción realizada, sino también a llamar la atención sobre la necesidad de evaluar empíricamente la calidad de dicha traducción. Además, un análisis de la equivalencia estructural no es suficiente, se requiere también llevar a cabo un análisis de la equivalencia de los items para detectar aquéllos que presentan un funcionamiento diferencial en las diferentes versiones del instrumento.

Como ya hemos comentado, en este estudio se han detectado un alto porcentaje de items que presentan FDI. El siguiente paso sería identificar las causas de este funcionamiento diferencial. Una primera aproximación, implica considerar problemas en la traducción realizada, es decir, que el FDI detectado sea debido a que el ítem no presenta exactamente el mismo significado en sus dos versiones. Conseguir items lingüísticamente equivalentes supone que estos presenten el mismo significado en los diferentes idiomas a los que es traducido, de tal modo que estas versiones del mismo ítem representen estímulos que evoquen la misma respuesta en aquellos sujetos que presentan el mismo nivel en el rasgo medido, independientemente del idioma en el que se administre la escala. Según Richards (1953), "la traducción es probablemente el tipo de acontecimiento más complejo producido hasta el momento en la evolución del cosmos" (pag. 250). Esta afirmación parece un tanto exagerada, sin embargo, los resultados obtenidos en este trabajo parecen apoyar las palabras de este autor. El riguroso proceso de traducción llevado a cabo ha generado items que aparentemente son lingüísticamente equivalentes, pero que empíricamente han demostrado no serlo. Tampoco queremos dar una visión negativa y pesimista sobre las posibilidades de generar escalas equivalentes. De hecho, son varios

los estudios que ponen de manifiesto que la utilización de métodos de juicio en la traducción de cuestionarios permite obtener versiones equivalentes en otros idiomas (Hulin y Mayer, 1986, Reise y cols, 1993).

Además, la traducción inadecuada no es la única causa que puede explicar el FDI de los items. Otra explicación es que este funcionamiento diferencial refleje diferencias reales entre las culturas analizadas. Como ya hemos comentado en el capítulo 3, los métodos de detección de FDI permiten detectar el funcionamiento diferencial de los items, pero no ofrecen información que permita interpretar dichas diferencias. Por lo tanto, la identificación de las causas del FDI detectado en ciertos items, representaría otro aspecto en el que profundizar en futuras investigaciones.

Como conclusión final, señalar que las principales aportaciones de este estudio han sido, en primer lugar la elaboración de una versión en castellano de la escala de autoconcepto físico PSDQ siguiendo un proceso riguroso de traducción y adaptación; en segundo lugar confirmar la equivalencia estructural de las dos versiones del instrumento, lo que de forma adicional ofrece evidencia a favor de la validez de constructo del PSDQ; la obtención de un grupo de items invariantes, que señalan la posibilidad de realizar comparaciones transculturales del autoconcepto físico en base a la versión en castellano del PSDQ elaborada; y finalmente, ofrecer evidencia empírica de la utilidad del AFC con medias latentes para el estudio de la equivalencia de diferentes versiones instrumentos de medida.

Además, se han planteado cuestiones a resolver en futuras investigaciones. Por una parte, la posibilidad de realizar estudios de

análisis del FDI de los items del cuestionario, utilizando métodos diferentes de DFDI, lo que permitiría realizar comparaciones entre los resultados obtenidos con cada método. Por otra parte, la necesidad de desarrollar estudios cualitativos que permitan identificar las causas del FDI detectado y posibiliten realizar modificaciones con el objeto de aumentar el número de items invariantes (en aquellos casos en los que las causas detectadas sean problemas lingüísticos); o bien profundizar en el estudio de las diferencias del autoconcepto físico en los adolescentes australianos y españoles (en aquellos casos en los que el FDI refleje diferencias reales entre los dos grupos).

Finalmente, también somos conscientes de las limitaciones de este estudio. Por ejemplo, el hecho de que las muestras utilizadas no sean representativas de las poblaciones de adolescentes australianos y españoles limita la generalizabilidad de los resultados. Por otra parte, también somos conscientes de la interpretación realizada en el contexto del AFC con medias latentes respecto al parámetro de dificultad del ítem (o parámetro de evocación), requiere un modelo de respuesta continua, y que en este trabajo se ha utilizado una escala de respuesta graduada. Sin embargo, como ya ha sido justificado anteriormente, la escala continua puede ser considerada como un caso extremo de respuesta graduada, por lo que ésta representaría una aproximación adecuada de la primera.

APENDICE

CUESTIONARIO DE AUTOCONCEPTO FISICO

Nombre: _____ Edad: _____ Sexo: ☐ Hombre ☐ Mujer

Colegio: _____ Curso: _____ Localidad: _____

Aquí tienes una oportunidad para reflexionar sobre ti mismo/A. **ESTO NO ES UN EXAMEN.** No hay respuestas correctas o incorrectas, ni respuestas que sean mejores que otras. Cada persona puede responder de forma diferente. Asegurate de que tus respuestas muestran lo que piensas sobre ti mismo/a. **POR FAVOR, NO COMENTES TUS RESPUESTAS CON NADIE.** Tus respuestas serán confidenciales.

El propósito de este estudio es analizar cómo se describen las personas físicamente. En las páginas siguientes te pediremos que pienses sobre algunas de tus características físicas: por ejemplo, si eres guapo/a, si eres fuerte, si se te dan bien los deportes, si haces ejercicio de forma regular, si tus movimientos son coordinados, si te pones enfermo/a muy a menudo, etcétera. Responde a cada frase rápidamente, tal y como te sientes ahora. Por favor, no dejes ninguna frase sin contestar.

Cuando estés listo/a para empezar, lee cada frase y decide tu respuesta. Hay seis respuestas posibles para cada frase: "Totalmente Verdadero", "Totalmente Falso", y cuatro respuestas intermedias. Junto a cada frase encontraras seis números, uno para cada una de las posibles respuestas. Las respuestas están escritas encabezando cada una de las columnas de números. Elige tu respuesta a la frase y rodea con un círculo (O) el número que esta bajo la respuesta elegida. Por favor, NO digas tu respuesta en voz alta ni la comentes con nadie.

Antes de comenzar te vamos a poner tres ejemplos. Yo he contestado dos de las tres frases para mostrarte cómo hacerlo. En la tercera debes elegir tu propia respuesta y rodearla con un círculo (O).

	Totalmente Falso	Bastante falso	Más falso que verdadero	Más verdadero que falso	Bastante verdadero	Totalmente Verdadero
1. Me gusta leer tebeos	1	2	3	4	5	(6)

(He rodeado con un círculo el número 6 que está debajo de la respuesta "Totalmente Verdadero". Esto significa que realmente me gusta leer tebeos. Si no me gustara demasiado leer tebeos habría respondido 1 ("Totalmente Falso") o 2 ("Bastante falso").

2. En general soy limpio y ordenado	1	2	(3)	4	5	6

(He contestado "Más falso que verdadero" porque realmente no soy muy ordenado, pero tampoco soy completamente desordenado).

3. Me gusta ver la televisión	1	2	3	4	5	6

(En esta frase tienes que elegir la respuesta que mejor te represente. Primero debes decidir si la frase es "Totalmente Verdadera", "Totalmente Falsa", u otra respuesta intermedia. Si realmente te gusta mucho ver la televisión, deberías contestar "Totalmente Verdadero" poniendo un círculo alrededor del último número ((6)). Si odias ver la televisión, deberías responder "Totalmente Falso" poniendo un círculo alrededor del primer número ((1)) Si no te gusta demasiado la televisión pero la ves de vez en cuando, podrías decidirte por marcar el 2 ("Bastante falso") o el 3 ("Más falso que verdadero").

Si mientras contestas las frases encuentras alguna palabra que no entiendes, levanta la mano, y pregúnta a la persona que te ha entregado el cuestionario. Si quieres cambiar una respuesta que ya has señalado, debes tachar el círculo con una cruz y poner un nuevo círculo alrededor de otro número de la misma línea. Asegúrate en todas las frases de que el numero que señalas se encuentra en la misma línea que la frase que estás contestando. Debes indicar solamente una respuesta por frase. No dejes ninguna frase sin contestar, incluso cuando no estés seguro/a de qué número rodear.

Por favor, si tienes ahora alguna pregunta, levanta la mano. Si no, ya puedes comenzar .

	Totalmente Falso	Bastante Falso	Más falso que Verdadero	Más verdadero que Falso	Bastante Verdadero	Totalmente Verdadero
1. Cuando estoy enfermo/a, me encuentro tan mal que no puedo ni levantarme de la cama	1	2	3	4	5	6
2. Me siento seguro/a realizando movimientos que requieren coordinación	1	2	3	4	5	6
3. Varias veces a la semana realizo ejercicios o deportes lo suficientemente duros como para hacerme respirar fuerte	1	2	3	4	5	6
4. Estoy demasiado gordo/a	1	2	3	4	5	6
5. La gente piensa que soy bueno/a en los deportes	1	2	3	4	5	6
6. Fisicamente, estoy satisfecho/a con el tipo de persona que soy	1	2	3	4	5	6
7. Teniendo en cuenta mi edad, soy atractivo/a	1	2	3	4	5	6
8. Soy una persona físicamente fuerte	1	2	3	4	5	6
9. Soy bastante bueno/a doblándome y retorciendo mi cuerpo	1	2	3	4	5	6
10. Puedo correr largas distancias sin parar	1	2	3	4	5	6
11. En general, la mayoría de las cosas que hago me salen bien	1	2	3	4	5	6
12. Normalmente cojo todas las enfermedades (gripe, virus, resfriados, etc.) que hay por ahí	1	2	3	4	5	6
13. Me resulta fácil controlar los movimientos de mi cuerpo	1	2	3	4	5	6
14. Suelo hacer ejercicio o actividades que me hacen respirar fuerte	1	2	3	4	5	6
15. Mi cintura es demasiado ancha	1	2	3	4	5	6
16. Se me dan bien la mayoría de deportes	1	2	3	4	5	6
17. Físicamente, me siento contento/a conmigo mismo/a	1	2	3	4	5	6

	Totalmente Falso	Bastante Falso	Más falso que Verdadero	Más verdadero que Falso	Bastante Verdadero	Totalmente Verdadero
18. Tengo una cara agradable	1	2	3	4	5	6
19. Tengo mucha fuerza física	1	2	3	4	5	6
20. Mi cuerpo es flexible	1	2	3	4	5	6
21. Obtendría buenos resultados en una prueba de resistencia física	1	2	3	4	5	6
22. No tengo mucho de lo que sentirme orgulloso/a	1	2	3	4	5	6
23. Estoy enfermo/a tan a menudo que no puedo hacer todas las cosas que quisiera	1	2	3	4	5	6
24. Soy bueno/a realizando movimientos que requieren coordinación	1	2	3	4	5	6
25. Tres o cuatro veces a la semana y al menos durante media hora, hago ejercicio o actividades que me hacen respirar fuerte	1	2	3	4	5	6
26. Tengo demasiada grasa en mi cuerpo	1	2	3	4	5	6
27. La mayoría de deportes me resultan fáciles	1	2	3	4	5	6
28. Me siento satisfecho/a con mi apariencia física y con lo que puedo hacer físicamente	1	2	3	4	5	6
29. Soy más guapo/a que la mayoría de mis amigos/as	1	2	3	4	5	6
30. Soy más fuerte que la mayoría de los chicos/as de mi edad	1	2	3	4	5	6
31. Mi cuerpo es rígido y nada flexible	1	2	3	4	5	6
32. Podría correr durante 5 kilómetros sin parar	1	2	3	4	5	6
33. Siento que mi vida no es demasiado útil	1	2	3	4	5	6
34. Casi nunca me pongo enfermo/a	1	2	3	4	5	6
35. En la mayoría de las actividades físicas, puedo realizar los movimientos con armonía	1	2	3	4	5	6
36. Hago actividades físicas (como correr, bailar, ir en bici, aerobic, gimnasia o nadar) por lo menos tres veces a la semana	1	2	3	4	5	6
37. Peso demasiado	1	2	3	4	5	6
38. Tengo buenas habilidades deportivas	1	2	3	4	5	6
39. Físicamente, me siento satisfecho/a conmigo mismo/a	1	2	3	4	5	6
40. Soy feo/a	1	2	3	4	5	6
41. Soy débil y casi no tengo músculo	1	2	3	4	5	6
42. Puedo doblar y mover bien las diversas partes de mi cuerpo en la mayoría de las direcciones	1	2	3	4	5	6
43. Creo que podría correr una distancia larga sin cansarme	1	2	3	4	5	6
44. En general, no valgo para nada	1	2	3	4	5	6
45. Me pongo enfermo/a con mucha frecuencia	1	2	3	4	5	6
46. Realizo con facilidad movimientos que requieren coordinación	1	2	3	4	5	6
47. Practico muchos deportes, baile, gimnasia u otras actividades físicas	1	2	3	4	5	6
48. Mi barriga es demasiado grande	1	2	3	4	5	6
49. Se me dan mejor los deportes que a la mayoría de mis amigos/as	1	2	3	4	5	6
50. Me siento satisfecho/a con quien soy y con lo que puedo hacer físicamente	1	2	3	4	5	6
51. Soy guapo/a	1	2	3	4	5	6
52. Obtendría buenos resultados en una prueba de fuerza	1	2	3	4	5	6
53. Creo que tengo bastante flexibilidad para la práctica de la mayoría de los deportes	1	2	3	4	5	6
54. Puedo mantenerme físicamente activo/a durante un periodo largo de tiempo sin cansarme	1	2	3	4	5	6
55. Hago bien la mayoría de las cosas que hago	1	2	3	4	5	6
56. Cuando me pongo enfermo/a me cuesta mucho tiempo recuperarme	1	2	3	4	5	6
57. Me muevo con gracia y coordinación cuando practico deportes y actividades	1	2	3	4	5	6
58. Practico deportes, ejercicio, baile u otras actividades físicas casi todos los días	1	2	3	4	5	6
59. La gente piensa que estoy gordo/a	1	2	3	4	5	6
60. Juego bien en los deportes	1	2	3	4	5	6
61. Estoy satisfecho/a con cómo soy físicamente	1	2	3	4	5	6
62. Nadie piensa que soy guapo/a	1	2	3	4	5	6
63. Se me da bien levantar objetos pesados	1	2	3	4	5	6
64. Creo que obtendría buenos resultados en una prueba de flexibilidad	1	2	3	4	5	6
65. Se me dan bien las actividades de resistencia física, como las carreras de larga distancia, el aerobic, el ciclismo o la natación	1	2	3	4	5	6
66. En general, tengo mucho de lo que sentirme orgulloso/a	1	2	3	4	5	6
67. Me pongo enfermo/a y tengo que ir al médico con más frecuencia que la mayoría de los chicos/as de mi edad	1	2	3	4	5	6
68. En general, soy un fracaso	1	2	3	4	5	6
69. Normalmente me mantengo sano/a, incluso cuando mis amigos/as se ponen enfermos	1	2	3	4	5	6
70. Nada de lo que hago parece salir bien	1	2	3	4	5	6

CUESTIONARIO DE AUTOCONCEPTO FISICO

Nombre: _____ Edad: _____ Sexo: ☐ Hombre ☐ Mujer

Colegio: _____ Curso: _____ Localidad: _____

Aquí tienes una oportunidad para reflexionar sobre ti mismo/A. **ESTO NO ES UN EXAMEN.** No hay respuestas correctas o incorrectas, ni respuestas que sean mejores que otras. Cada persona puede responder de forma diferente. Asegurate de que tus respuestas muestran lo que piensas sobre ti mismo/a. **POR FAVOR, NO COMENTES TUS RESPUESTAS CON NADIE.** Tus respuestas serán confidenciales.

El propósito de este estudio es analizar cómo se describen las personas físicamente. En las páginas siguientes te pediremos que pienses sobre algunas de tus características físicas: por ejemplo, si eres guapo/a, si eres fuerte, si se te dan bien los deportes, si haces ejercicio de forma regular, si tus movimientos son coordinados, si te pones enfermo/a muy a menudo, etcétera. Responde a cada frase rápidamente, tal y como te sientes ahora. Por favor, no dejes ninguna frase sin contestar.

Cuando estés listo/a para empezar, lee cada frase y decide tu respuesta. Hay seis respuestas posibles para cada frase: "Totalmente Verdadero", "Totalmente Falso", y cuatro respuestas intermedias. Junto a cada frase encontraras seis números, uno para cada una de las posibles respuestas. Las respuestas están escritas encabezando cada una de las columnas de números. Elige tu respuesta a la frase y rodea con un círculo (O) el número que esta bajo la respuesta elegida. Por favor, NO digas tu respuesta en voz alta ni la comentes con nadie.

Antes de comenzar te vamos a poner tres ejemplos. Yo he contestado dos de las tres frases para mostrarte cómo hacerlo. En la tercera debes elegir tu propia respuesta y rodearla con un círculo (O).

	Totalmente Falso	Bastante falso	Más falso que verdadero	Más verdadero que falso	Bastante verdadero	Totalmente Verdadero
1. Me gusta leer tebeos	1	2	3	4	5	(6)

(He rodeado con un círculo el número 6 que está debajo de la respuesta "Totalmente Verdadero". Esto significa que realmente me gusta leer tebeos. Si no me gustara demasiado leer tebeos habría respondido 1 ("Totalmente Falso") o 2 ("Bastante falso").

2. En general soy limpio y ordenado	1	2	(3)	4	5	6

(He contestado "Más falso que verdadero" porque realmente no soy muy ordenado, pero tampoco soy completamente desordenado).

3. Me gusta ver la televisión	1	2	3	4	5	6

(En esta frase tienes que elegir la respuesta que mejor te represente. Primero debes decidir si la frase es "Totalmente Verdadera", "Totalmente Falsa", u otra respuesta intermedia. Si realmente te gusta mucho ver la televisión, deberías contestar "Totalmente Verdadero" poniendo un círculo alrededor del último número ((6)). Si odias ver la televisión, deberías responder "Totalmente Falso" poniendo un círculo alrededor del primer número ((1)) Si no te gusta demasiado la televisión pero la ves de vez en cuando, podrías decidirte por marcar el 2 ("Bastante falso") o el 3 ("Más falso que verdadero").

Si mientras contestas las frases encuentras alguna palabra que no entiendes, levanta la mano, y pregúnta a la persona que te ha entregado el cuestionario. Si quieres cambiar una respuesta que ya has señalado, debes tachar el círculo con una cruz y poner un nuevo círculo alrededor de otro número de la misma línea. Asegúrate en todas las frases de que el numero que señalas se encuentra en la misma línea que la frase que estás contestando. Debes indicar solamente una respuesta por frase. No dejes ninguna frase sin contestar, incluso cuando no estés seguro/a de qué número rodear.

Por favor, si tienes ahora alguna pregunta, levanta la mano. Si no, ya puedes comenzar .

	Totalmente Falso	Bastante Falso	Más falso que Verdadero	Más verdadero que Falso	Bastante Verdadero	Totalmente Verdadero
1. Cuando estoy enfermo/a, me encuentro tan mal que no puedo ni levantarme de la cama	1	2	3	4	5	6
2. Me siento seguro/a realizando movimientos que requieren coordinación	1	2	3	4	5	6
3. Varias veces a la semana realizo ejercicios o deportes lo suficientemente duros como para hacerme respirar fuerte	1	2	3	4	5	6
4. Estoy demasiado gordo/a	1	2	3	4	5	6
5. La gente piensa que soy bueno/a en los deportes	1	2	3	4	5	6
6. Físicamente, estoy satisfecho/a con el tipo de persona que soy	1	2	3	4	5	6
7. Teniendo en cuenta mi edad, soy atractivo/a	1	2	3	4	5	6
8. Soy una persona físicamente fuerte	1	2	3	4	5	6
9. Soy bastante bueno/a doblándome y retorciendo mi cuerpo	1	2	3	4	5	6
10. Puedo correr largas distancias sin parar	1	2	3	4	5	6
11. En general, la mayoría de las cosas que hago me salen bien	1	2	3	4	5	6
12. Normalmente cojo todas las enfermedades (gripe, virus, resfriados, etc.) que hay por ahí	1	2	3	4	5	6
13. Me resulta fácil controlar los movimientos de mi cuerpo	1	2	3	4	5	6
14. Suelo hacer ejercicio o actividades que me hacen respirar fuerte	1	2	3	4	5	6
15. Mi cintura es demasiado ancha	1	2	3	4	5	6
16. Se me dan bien la mayoría de deportes	1	2	3	4	5	6
17. Físicamente, me siento contento/a conmigo mismo/a	1	2	3	4	5	6

	Totalmente Falso	Bastante Falso	Más falso que Verdadero	Más verdadero que Falso	Bastante Verdadero	Totalmente Verdadero
18. Tengo una cara agradable	1	2	3	4	5	6
19. Tengo mucha fuerza física	1	2	3	4	5	6
20. Mi cuerpo es flexible	1	2	3	4	5	6
21. Obtendría buenos resultados en una prueba de resistencia física	1	2	3	4	5	6
22. No tengo mucho de lo que sentirme orgulloso/a	1	2	3	4	5	6
23. Estoy enfermo/a tan a menudo que no puedo hacer todas las cosas que quisiera	1	2	3	4	5	6
24. Soy bueno/a realizando movimientos que requieren coordinación	1	2	3	4	5	6
25. Tres o cuatro veces a la semana y al menos durante media hora, hago ejercicio o actividades que me hacen respirar fuerte	1	2	3	4	5	6
26. Tengo demasiada grasa en mi cuerpo	1	2	3	4	5	6
27. La mayoría de deportes me resultan fáciles	1	2	3	4	5	6
28. Me siento satisfecho/a con mi apariencia física y con lo que puedo hacer físicamente	1	2	3	4	5	6
29. Soy más guapo/a que la mayoría de mis amigos/as	1	2	3	4	5	6
30. Soy más fuerte que la mayoría de los chicos/as de mi edad	1	2	3	4	5	6
31. Mi cuerpo es rígido y nada flexible	1	2	3	4	5	6
32. Podría correr durante 5 kilómetros sin parar	1	2	3	4	5	6
33. Siento que mi vida no es demasiado útil	1	2	3	4	5	6
34. Casi nunca me pongo enfermo/a	1	2	3	4	5	6
35. En la mayoría de las actividades físicas, puedo realizar los movimientos con armonía	1	2	3	4	5	6
36. Hago actividades físicas (como correr, bailar, ir en bici, aerobic, gimnasia o nadar) por lo menos tres veces a la semana	1	2	3	4	5	6
37. Peso demasiado	1	2	3	4	5	6
38. Tengo buenas habilidades deportivas	1	2	3	4	5	6
39. Físicamente, me siento satisfecho/a conmigo mismo/a	1	2	3	4	5	6
40. Soy feo/a	1	2	3	4	5	6
41. Soy débil y casi no tengo músculo	1	2	3	4	5	6
42. Puedo doblar y mover bien las diversas partes de mi cuerpo en la mayoría de las direcciones	1	2	3	4	5	6
43. Creo que podría correr una distancia larga sin cansarme	1	2	3	4	5	6
44. En general, no valgo para nada	1	2	3	4	5	6
45. Me pongo enfermo/a con mucha frecuencia	1	2	3	4	5	6
46. Realizo con facilidad movimientos que requieren coordinación	1	2	3	4	5	6
47. Practico muchos deportes, baile, gimnasia u otras actividades físicas	1	2	3	4	5	6
48. Mi barriga es demasiado grande	1	2	3	4	5	6
49. Se me dan mejor los deportes que a la mayoría de mis amigos/as	1	2	3	4	5	6
50. Me siento satisfecho/a con quien soy y con lo que puedo hacer físicamente	1	2	3	4	5	6
51. Soy guapo/a	1	2	3	4	5	6
52. Obtendría buenos resultados en una prueba de fuerza	1	2	3	4	5	6
53. Creo que tengo bastante flexibilidad para la práctica de la mayoría de los deportes	1	2	3	4	5	6
54. Puedo mantenerme físicamente activo/a durante un periodo largo de tiempo sin cansarme	1	2	3	4	5	6
55. Hago bien la mayoría de las cosas que hago	1	2	3	4	5	6
56. Cuando me pongo enfermo/a me cuesta mucho tiempo recuperarme	1	2	3	4	5	6
57. Me muevo con gracia y coordinación cuando practico deportes y actividades	1	2	3	4	5	6
58. Practico deportes, ejercicio, baile u otras actividades físicas casi todos los días	1	2	3	4	5	6
59. La gente piensa que estoy gordo/a	1	2	3	4	5	6
60. Juego bien en los deportes	1	2	3	4	5	6
61. Estoy satisfecho/a con cómo soy físicamente	1	2	3	4	5	6
62. Nadie piensa que soy guapo/a	1	2	3	4	5	6
63. Se me da bien levantar objetos pesados	1	2	3	4	5	6
64. Creo que obtendría buenos resultados en una prueba de flexibilidad	1	2	3	4	5	6
65. Se me dan bien las actividades de resistencia física, como las carreras de larga distancia, el aerobic, el ciclismo o la natación	1	2	3	4	5	6
66. En general, tengo mucho de lo que sentirme orgulloso/a	1	2	3	4	5	6
67. Me pongo enfermo/a y tengo que ir al médico con más frecuencia que la mayoría de los chicos/as de mi edad	1	2	3	4	5	6
68. En general, soy un fracaso	1	2	3	4	5	6
69. Normalmente me mantengo sano/a, incluso cuando mis amigos/as se ponen enfermos	1	2	3	4	5	6
70. Nada de lo que hago parece salir bien	1	2	3	4	5	6

BIBLIOGRAFIA

Abraham, S.F., y Beaumont, P.J. (1982). How patients describe bulimia or binge-eating. Psychological Medicine, 12, 625-635.

Abu-Lughod, L. (1985). Honor and the sentiments of loss in a Bedouin society. American Ethnology, 12, 245-261.

Abu-Lughod, L. (1986). Veiled Sentiments: Honor and Poetry in a Bedouin Society. Berkeley: University California Press. 317 pp.

Adams, G.R. (1977). Physical attractiveness, personality, and social reactions to peer pressure. Journal of Psychology, 96, 287-296.

Adler, N.J. (1991). International Dimensions of Organizational Behavior. Boston: PWS-Kent. 2ª ed.

Alwin, D.F. y Jackson, D.J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. En D.J. Jackson, & E.F. Borgotta (Eds.), Factor analysis and measurement in sociological research: A multidimensional perspective. Beverly Hills, CA: Sage.

Amir, Y. y Sharon, I. (1987). Are social psychological laws cross-culturally valid? Journal of Cross-Cultural Psychology, 18, 383-470.

Amirkhan, J.H. (1990). A factor-analytically derived measure of coping: The Coping Strategy Indicator. Journal of Personality and Social Psychology, 59, 1066-1074.

Amon, J. (1987). Estadística para psicólogos 2. Madrid: Pirámide.

Anastasi, A. (1976). Psychological testing (4ª ed.). New York: Macmillan.

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. En P.W. Holland & H. Wainer (Eds.), Differential item functioning (p.3-23). Hillsdale, NJ: Lawrence Erlbaum.

Bakker, F.C. (1988). Personality differences between young dancers and nondancers. Personality and Individual Differences, 9, 121-131.

Ben-Porath, Y.S.; Hoffman-Chemi, M.A.A.; y Tellegen, A. (1995). A cross-cultural study of personality with the multidimensional personality questionnaire. Journal of Cross-Cultural Psychology, 26 (4), 360-373.

Bentler, P.M. (1988). Theory and implementation of EQS: A structural equations program. Los Angeles: BMDP Statistical Software, Inc.

Bentler, P.M. y Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Berk, R.A. (ed.) (1982). Handbook of methods for detecting item bias. Baltimore: Johns Hopkins University Press.

Berry, J.W. (1967). Independence and conformity in subsistence-level societies. Journal of Personality and Social Psychology, 7, 415-418.

Berry, J.W. (1969). On cross-cultural comparability. International Journal of Psycholgy, 4, 119-128.

Berry, J.W. (1976). Human ecology and cognitive style: Comparative studies in cultural and psychological adaptation. Beverly Hills, CA: Sage.

Berry, J.W.; Poortinga, Y.H.; Pandey, J. (eds.) (1996). Handbook of Cross-Cultural Psychology. Second Edition (3 vols.). Needham, MA: Allyn & Bacon.

Berry, J.W.; Poortinga, Y.H.; Segall, M.H.; y Dasen, P.R. (1992). Cross-Cultural Psychology: Research and Applications. Cambridge: Cambridge University Press.

Berry, J.W.; Van de Koppel, J.M.H.; Sénéchal, C.; Annis, R.C.; Bahuchet, S.; Cavalli-Sforza, L.L.; y Witkin, H.A. (1986). On the edge of the forest: Cultural adaptation and cognitive development in Central Africa. Lisse: Swets & Zeitlinger.

Best, D.L. y Williams, J.E. (1994). Masculinity/feminity in the self and ideal self-descriptions of university students in fourteen countries. Ver Bouvy et al. 1994, pp. 297-306.

Binet, A. y Simon, T. (1973). The development of intelligence in children. New York: Arno. (Trabajo original publicado en 1916).

Boersma, F.J., y Chapman, J.W. (1992). Perception of Ability Scale for Students. Los Angeles, CA: Western Psychological Services.

Bollen, K.A. (1989). Structural equations with latent variables. New York: Wiley.

Bond, M.H. (1991). Chinese values and health: A cross-cultural examination. Psychology and Health, 5, 137-152.

Bond, M.H. y Smith P.B. (1996). Cross-cultural social and organizational psychology. Annual Review of Psychology, 47, 205-235.

Bouvy, A.M.; Van de Vijver, F.J.R.; Boski, P.; Schmitz, P. (eds.) (1994). Journeys into Cross-Cultural Psychology. Amsterdam: Swets & Zeitlinger.

Boyle, G.J. (1994). Self-Description Questionnaire II: A review. Test Critiques, 10, 632-643.

Bracken, B.A. (1992). Multidimensional Self Concept Scale. Austin, TX: Pro-Ed.

Briggs, J.L. (1970). Never in Anger: Portrait of an Eskimo Family. Cambridge, Mass.: Harvard University Press. 379 pp.

Brislin, R. (1993). Understanding Culture's Influence on Behavior. Fort Worth, TX: Harcourt, Brace, Jovanovich.

Brislin, R.W. (1980). Translation and content analysis of oral and written material. En H. C. Triandis y J.W. Berry (eds.), Handbook of Cross-Cultural Psychology (Vol. 1, pp. 389-444). Boston: Allyn and Bacon.

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 Research (pp. 137-164). Newbury Park, CA: Sage.

Brislin, R.W.; Lonner, W.J. y Thorndike, R. (1973). Cross-cultural research methods. New York: Wiley.

Brown, L., y Alexander, J. (1991). Self-Esteem Index. Austin, TX: Pro-Ed.

Brown, R. y Marcoulides, G.A. (1996). A cross-cultural comparison of the Brown Locus of Control Scale. Educational and Psychological Measurement, 56 (5), 858-863.

Browne, M.W. (1990). MUTMUM PC: User's guide. Columbus: Ohio State University, Department of Psychology.

Burns, R.B. (1979). The self-concept: Theory, measurement, development and behaviour. London: Longman.

Buss, D.M., Abbott, M., Angleitner, A., Asherian, A., Biaggio, A., Blanco-Villaseñor, A., Bruchon-Schweitzer, M., Chu, H., Czapinski, J., De Raad, B., Ekehammar, B., El Lohamy, N., Fioravanti, M., Georgas, J., Gerde, P., Guttman, R., Hazan, F., Iwawaki, S., Janakiramaiah, N., Khosrokhani, F., Kretner, S., Lachenicht, L., Lee, M., Liik, M., Little, B., Mika, S., Moadel-Shadid, M., Moane, G., Montero, M., Mundy-Castle, A.C., Niit, T., Nsenduluka, E., Pienkowski, R., Pirttila-Blackman, A.M., Ponce de Leon, J., Rousseau, J., Runco, M.A., Safir, M.P., Samuels, C., Sanitioso, R., Serpell, R., Smind, N., Spencer, C., Tadinac, M., Todorova, E.N., Troland, K., Van den Brande, L., Van Heck, G., Van Langenhove L., y Yang, K.S. (1990). International preferences in selecting mate. A study of 37 cultures. Journal of Cross-Cultural Psychology, 21, 5-47.

Byrne, B.M. (1984). The general/academic self-concept nomological network: A review of construc validation research. Review of Educational Research, 54, 427-456.

Byrne, B.M. (1989). A primer of LISREL: Basic applications and programming for confirmatory factor analytic models. New York: Springer-Verlag.

Byrne, B.M., Shavelson, R.J., y Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measuremt invariance. Psychological bulletin, 105, 456-466.

Camilli, G. (1993). The case against DIF techniques based on internal criteria: Do item bias procedures obscure test fairness issues? En P.W. Holland & H. Wainer (Eds.). Differential item functioning: Theory and practice. Hillsdale, NJ: Lawrence Erlbaum.

Camilli, G. y Shepard, L.A. (1987). The inadequacy of ANOVA for detecting test bias. Journal of Educational Statistics, 12, 87-99.

Camilli, G. y Shepard, L.A. (1994). Methods for identifying biased test items. Thousand Oaks, CA: Sage Publications.

Campbell, D.T. y Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-mutimethod matrix. Psychological Bulletin, 56, 81-105.

Candell, G.L. y Hulin, C.L. (1986). Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. Journal of Cross-Cultural Psychology, 17 (4), 417-440.

Cattell, R.B. (1940). A culture-free intelligence test, I. Journal of Educational Psychology, 31, 176-199.

Cattell, R.B. y Cattell, A.K.S. (1963). Culture Fair Intelligence Test. Champaing, IL: Institute for Personality and Ability Testing.

Cleary, T.A. y Hilton, T.L. (1968). An investigation of item bias. Educational and Psychological Measurement, 28, 61-75.

Clifford, E. (1971). Body satisfaction in adolescence. Perceptual and Motor Skills, 33, 119-125.

Cole, D.A. y Maxwell, S.E. (1985). Multitrait-multimethod comparisons across populations: A confirmatory factor analytic approach. Multivariate Behavioral Research, 20, 389-417.

Cole, M. (1988). Cross-cultural research in the sociohistorical tradition. Human Development, 31, 137-157.

Cole, M. (1990). Cultural psychology: a once and future discipline? En Cross-cultural Perspectives. Nebraska Symposium on Motivation, 1989, ed. J.J. Berman. Lincoln: University of Nebraska Press. 227 pp.

Collins, M.E. (1991). Body figure perceptions and preferences among preadolescent children. International Journal of Eating Disorders, 10, 199-208.

Cook, T.D. y Campbell, D.T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally.

Cooley, C.H. (1902). Human nature and the social order. New York: Scribner's.

Coopersmith, S.A. (1967). The antecedents of self-esteem. Palo Alto, CA: Consulting Psychologists Press, Inc.

Coopersmith, S.A. (1981). Coopersmith Self-Esteem Inventory. Palo Alto, CA: Consulting Psychologists Press, Inc.

Cornell, D.G., Pelton, G.M., Bassin, L.E., Landrum, M., Ramsay, S.G., Cooley, M.R., Lynch, K.A., y Hamrick, E. (1990). Self-concept and peer status among gifted program youth. Journal of Educational Psychology, 82, 456-463.

Crapanzano, V. (1980). Tuhami: Portrait of a Moroccan. University Chicago Press. 187 pp.

Cziko, G. (1987). Review of the bilingual Syntax Measure I. En J.C. Alderson y K.J. Krahnke (eds.), Reviews of English Language Proficiency Tests. Washington, DC: TESOL.

Cheung, F.M.; Leung, K.; Fan, R.M.; Song, W.Z.; Zhang, J.X.; y Zhang, J. P. (1996). Development of the Chinesse Personality Assessment Inventory. Journal of Cross-Cultural Psychology, 27 (2), 181-199.

Drasgow, F. y Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. Journal of Applied Psychology, 70 (4), 662-680.

Drasgow, F. y Parsons, C.K. (1983). Application of unidimensional item response theory models to multidimensional data. Applied Psychological Measurement, 7, 189-199.

Dyer, C.O. (1964). Construct validity of self-concept by a multitrait-multimethod analysis (Tesis doctoral, Universidad de Michigan, 1963). Dissertation Abstracts International, 25, 8154. (University Microfilms N° 64-8154).

Earley, C. (1989). Social loafing and collectivism: A comparison of the United States and the People's Republic of China. Administrative Science Quarterly, 34, 565-581.

Ellis, B.B. (1989). Differential item functioning: Implications for test translations. Journal of Applied Psychology, 74, 912-921.

Ellis, B.B. (1991). Item response theory: A tool for assessing the equivalence of translated tests. Bulletin of the Internatioanl Test Commission, 18, 33-51.

Ellis, B.B. y Kimmel, H.D. (1992). Identification of unique cultural response patterns by means of item response theory. Journal of Applied Psychology, 77, 177-184.

Ellis, B.B.; Kimmel, H.D.; Diaz Guerrero, R.; Cana, J.; y Bajo, M.T. (1995). Love and power in Mexico, Spain and the United States. Journal of Cross-Cultural Psychology, 25, 525-540.

Ellis, B.B; Becker, P.; y Kimmel, H. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). Journal of Cross-Cultural Psychology, 24 (2), 133-148.

Engelhard, G.; Hansche, L.; y Rutledge, K.E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. Applied Measurement in Education, 3 (4), 347-360.

Erez, M. y Earley, P.C. (1993). Culture, Self-Identity and Work. New York: Oxford University Press.

Everson, H.T., Millsap, R.E., y Rodriguez, C.M. (1991). Isolating gender differences in test anxiety: A confirmatory factor analysis of the test anxiety inventory. Educational and Psychological Measurement, 51, 243-251.

Ferrando, P.J. (1996a). Relaciones entre el análisis factorial y la teoría de respuesta a los items. En J. Muñiz (coor.), Psicometría (p. 555-612). Madrid: Universitas.

Ferrando, P.J. (1996b). Calibration of invariant item parameters in a continuous item response model using the extended Lisrel measurement submodel. Multivariate Behavioral Research, 31 (4), 419-439.

Fidalgo, A.M. (1996). Funcionamiento diferencial de los items. Procedimiento Mantel-Haenszel y modelos loglineales. Tesis doctoral no publicada. Universidad de Oviedo.

Fidalgo, A.M. y Mellenbergh, G.J. (1995). Evaluación del procedimiento Mantel-Haenszel frente al método logit iterativo en la detección del funcionamiento diferencial de los items uniforme y no uniforme. Comunicación presentada al IV Symposium de Metodología de las Ciencias del Comportamiento, La Manga del Mar Menor.

Fisher, S., y Cleveland, S.E. (1968). Body image and personality. (rev.ed.). New York: Dover Publications.

Fitts, W.H. (1965). Tennessee Self-Concept Scale. Nashville, TN: Counselor Recordings and Tests.

Fleming, J.S., y Courtney, B.E. (1984). The dimensionality of self-esteem, II: Hierarchical facet model for revised measurement scales. Journal of Personality and Social Psychology, 46, 404-421.

Fleming, J.S., y Watts, W.A. (1980). The dimensionality of self-esteem: Some results for a college sample. Journal of Personality and Social Pscyhology, 39, 921-929.

Fox, K.R. (1990). The Physical Self-Perception Profile manual. DeKalb, IL: Office for Health Promotion, Northern Illinois University.

Fox, K.R., y Corbin, C.B. (1989). The Physical Self-Perception Profile: Development and preliminary validation. Journal of Sport and Exercise Psychology, 11, 408-430.

Franzoi, S.L., y Shields, S.A. (1984). The Body Esteem Scale: Multidimensional structure and sex differences in a college population. Journal of Personality Assessment, 48, 173-178.

Frijda, N.H. y Mesquita, B. (1994). The social roles and functions of emotions. Ver Kitayama y Markus, 1994, pp. 51-87.

Garner, D.M., Garfinkel, P.E., Schwartz, D., y Thompson, M. (1980). Cultural expectations of thinness in women. Psychological Reports, 47, 483-491.

Geertz, C. (1984). Anti anti-relativism. American Anthropology, 86, 263-278.

Gómez Benito, J. (1996). Aportaciones de los modelos de estructuras de covariancia al análisis psicométrico. En J. Muñiz (coor.), Psicometría (p. 457-554). Madrid: Universitas.

Guida, F.V. y Ludlow, L.H. (1989). A cross-cultural study of test anxiety. Journal of Cross-Cultural Psychology, 20, 178-190.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. Japanese Psychological Research, 22(3), 144-149.

Hambleton, R.K. (1993). Translating achievement tests for use in cross-national studies. European Journal of 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, 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 (coor.), Psicometría (pp. 208-238). Madrid: Universitas.

Hambleton, R.K. y Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. European Journal of Psychological Assessment, 11 (3), 147-157.

Hambleton, R.K. y Rogers, H.J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2 (4), 313-334.

Hambleton, R.K. y Swaminathan, H. (1985). Item response theory: Principles and applications. Dordrecht: Kluwer-Nijhoff.

Hambleton, R.K.; Swaminathan, H.; y Rogers, H.J. (1991). Fundamentals of item response theory. Newbury, CA: Sage.

Harrison, D.A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. Journal of Educational Statistics, 11, 91-115.

Harter, S. (1982). The Perceived Competence Scale for Children. Child Development, 53, 87-97.

Harter, S. (1985). Self-Perception Profile for Children. Denver, CO: University of Denver Press.

Harter, S. (1986). Processes underlying the construction, maintenance and enchancement of self-concept in children. En J. Suls & A. Greenwald (Eds.), Psychological perspectives on the self (vol. 3, pags. 136-182). Hillsdale, NJ: Erlbaum.

Harvey, D.H.P., y Greenway, A.P. (1984). The self-concept of physically handicapped children and their nonhandicapped siblings: An empirical investigation. Journal of Child Psychology and Psychiatry, 25, 273-284.

Hatfield, E., y Sprecher, S. (1986). Mirror, mirror... The importance of looks in everyday life. Albany, NY: State University of New York Press.

Hattie, J. (1992). Self-concept. Hillsdale, NJ: Erlbaum.

Heelas, P.L.F. y Lock, A.J., eds. (1981). Indigenous Psychologies: The Anthropology of the Self. San Diego, California: Academic. 322 pp.

Holland, P.W. (1985). On the study of differential item performance without IRT. Procedings of the 27th Annual Conference of the Military Testing Association (Vol. 1, pp. 282-287). San Diego.

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

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

Hulin, C.L. (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. Journal of Cross-Cultural Psychology, 67, 115-142.

Hulin, Ch.L. y Mayer, L.J. (1986). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. Journal of Applied Psychology, 71 (1), 83-94.

Irvine, S.H. (1979). The place of factor analysis in cross-cultural methodology and its contribution to cognitive theory. En L. Eckensberger, W. Lonner, y Y.H. Poortinga (eds.), Cross-cultural contibutions to psychology (pp. 300-341). Lisse: Swets & Zeitlinger.

Irvine, S.H. y Carroll, W.K. (1980). Testing and assessment across cultures. En H.C. Triandis y J.W. Berry (Eds.), Handbook of cross-cultural psychology (Vol. 2, pp. 181-244). Boston: Allyn and Bacon.

Iwawaki, S.; Kashima, Y.; Leung, K. (eds.) (1992). Innovations in Cross-Cultural Psychology. Amsterdam: Swets & Zeitlinger.

Jackson, S., y Marsh, H.W. (1986). Athletic or antisocial: The female sport experience. Journal of Sport Psychology, 8, 198-211.

Jäger, A.O. (1963). Der Wilde-Test. Ein neues Intelligenzdiagnostikum. Zeitschrift für experimentelle und angewandte Psychologie, 10, 260-278.

Jäger, A.O., y Althoff, K. (1983). Der WILDE-Intelligenz-Test (WIT): Ein Strukturdiagnostikum. Göttingen, Federal Republic of Germany: Hogrefe.

Jahoda, G. (1992). Crossroads Between Culture and Mind: Continuities and Change in Theories of Human Nature. London: Harvester Wheatsheaf.

James, W. (1892). Psychology: The briefer course. New York: Henry Holt & Co.

James, W. (1963). The principles of psychology. New York: Holt, Rinehart & Winston. (Trabajo original publicado en 1890).

Jensen, A.R. (1969). How much can we boost IQ and scholastic achievement? Harvard Educational Review, 39, 1-123.

Jensen, A.R. (1980). Bias in mental testing. New York: Free Press.

Jöreskog, K.G. (1971). Simultaneous factor analysis in several populations. Psychometrika, 36, 409-426.

Jöreskog, K.G. y Sörbom, D. (1985). LISREL VI User's Guide. Mooresville, IN: Scientific Software Inc.

Jöreskog, K.G. y Sörbom, D. (1988). LISREL 7: A guide to the program and applications. Chicago: SPSS, Inc.

Jöreskog, K.G. y Sörbom, D. (1993). LISREL VIII: User's reference guide. Mooresville, IN: Scientific Software.

Jöreskog, K.G., y Sörbom, D. (1993). LISREL 8: Structural equation modeling with the SIMPLIS command language. Chicago: Scientific Software International.

Keats, D.M.; Munro, D.; y Mann, L. (1989). Heterogeneity in Cross-Cultural Psychology. Amsterdam: Swets & Zeitlinger.

Keeves, J.P. (1992). Learning science in a changing world: Cross-national studies of science achievement, 1970 to 1984. The Hague, The Netherlands: The International Association for the Evaluation of Educational Achievement.

Keith, L.K., y Bracken, B.A. (1996). Self-concept instrumentation: A historical and evaluative review. En B.A. Bracken (Eds.), Handbook of Self-Concept (pags. 91-170). New York: Wiley.

Keller, A., Ford, L.H., y Meacham, J.A. (1978). Dimensions of self-concept in preschool children. Developmental Psychology, 14, 483-489.

Kilbride, P.L. (1992). Anti anti-universalism: rethinking cultural psychology as anti anti-relativism. Rev. Anthropol. In press.

Kitayama, S. y Markus, H.R. (eds.) (1994). Emotion and Culture: Empirical Studies of Mutual Influence. Washington, DC: American Psychological Association.

Klesges, R.C., Haddock, C.K., Stein, R.S., Klesges, L.M., Eck, L.H., y Hanson, C.L. (1992). Relationship between psychosocial functioning and body fat in preschool children: A longitudinal investigation. Journal of Consulting and Clinical Psychology, 60, 793-796.

Knapp, L., y Knapp, R.R. (1976). Career Ability Placement Survey. San Diego, CA: EdITS.

Lanning, K. (1991). Consistency, scalability and personality measurement. New York: Springer-Verlag.

Lapointe, A.E.; Mead, N.A.; y Askew, J.M. (1992). Learning mathematics (Report N° 22-CAEP-01). Princeton, NJ: Educational Testing Service.

Lapointe, A.E.; Mead, N.A.; y Phillips, G.W. (1989). A world of differences: An international assessment of mathematics and science (Report N° 19-CAEP-01). Princeton, NJ: Educational Testing Service.

Lerner, R.M., Karabenick, S.A., y Stuart, J.L. (1973). Relations among physical attractiveness, body attitudes, and self-concept in male and female college students. Journal of Psychology, 85, 119-129.

Lerner, R.M., Orlos, J.B., y Knapp, J.R. (1976). Physical attractiveness, physical effectiveness, and self-concept on late adolescents. Adolescence, 11, 313-326.

Lerner, R.M., y Karabenick, S.A. (1974). Physical attractiveness, body attitudes, and self-concept in late adolescents. Journal of Youth and Adolescence, 3, 307-316.

Leung, K. (1987). Some determinants of reactions to procedural models for conflict resolution. Journal of Personality and Social Psychology, 53, 898-908.

Leung, K. y Drasgow, F. (1985). Relation between self-esteem and delinquent behavior in three ethnic groups: An application of item response theory. Journal of Cross-Cultural Psychology, 17, 151-167.

Levy, R.I. (1973). Tahitians: Mind and Experience in the Society Islands. University Chicago Press. 547 pp.

Levy, R.I. (1978). Tahitian gentleness and redundant controls. En Learning Non-Aggression, ed. A. Montagu, pp. 222-235. Oxford University Press.

Levy, R.I. (1983). Introduction: self and emotion. Ethos, 11, 128-134.

Lomax, R.G. (1985). A structural model of public and private schools. Journal of Experimental Education, 53, 216-226.

Lonner, W.J. y Malpass, R. (eds.) (1994). Psychology and Culture. Boston: Allyn & Bacon.

Lord, F.M. (1977). A study of item bias, using Item Characteristic Curve Theory. En Y.H. Poortinga (ed.), Basic problems in cross-cultural psychology (pp. 19-29). Lisse: Swets & Zeitlinger.

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

Loro, A.D., y Orleans, C.S. (1981). Binge eating in obesity: Preliminary findings and guidelines for behavioral analysis and treatment. Addictive Behaviors, 6, 155-166.

MacCallum, R.C. y Tucker, L.R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. Psychological Bulletin, 109, 501-511.

Maloney, M.J., McGuire, J.B., y Daniels, S.R. (1988). Reliability testing of a children's version of the Eating Attitude Test. Journal of the American Academy of Child and Adolescent Psychiatry, 27, 541-543.

Marsh, H.W. (1992). Self-Description Questionnaire II. Manual. Sydney: University of Western Sydney, Macarthur, Faculty of Education, Publication Unit.

Marsh, H.W. (1986). The Self-Description Questionnaire (SDQ): A theoretical and empirical basis for the measurement of multiple dimensions of preadolescent self-concept: A test manual and a research monograph. Faculty of Education, University of Sydney, NSW Australia.

Marsh, H.W. (1987a). The factorial invariance of responses by males and females to a multidimensional self-concept instrument: Substantive and methodological issues. Multivariate Behavioral Research, 22, 457-480.

Marsh, H.W. (1987b). The hierarchical structure of self-concept and the application of confirmatory hierarchical factors analysis. Journal of Educational Measurement, 24, 17-39.

Marsh, H.W. (1988). Self-Description Questionnaire, I. San Antonio, TX: the Psychological Corporation.

Marsh, H.W. (1990a). A multidimensional, hierarchical self-concept: Theoretical and empirical justification. Educational Psychology Review, 2, 77-172.

Marsh, H.W. (1990b). Self-Description Questionnaire, II. San Antonio, TX: the Psychological Corporation.

Marsh, H.W. (1990c). The structure of academic self-concept: The Marsh/Shavelson model. Journal of Educational Psychology, 82, 623-636.

Marsh, H.W. (1992) Self-Description Questionnaire (SDQ) III: A theoretical and empirical basis for the measurement of multiple dimensions of late adolescent self-concept: A test manual and a research monograph. Sydney: Faculty of Education, University of Western Sydney.

Marsh, H.W. (1993). The multidimensional structure of physical fitness: Invariance over gender and age. Research Quarterly for Exercise and Sport, 64, 256-273.

Marsh, H.W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. Structural equation modeling, 1, 5-34.

Marsh, H.W. (1996). Construct validity of Physical Self-Description Questionnaire responses: Relations to external criteria. Journal of Sport and Exercise Psychology, 18, 11-131.

Marsh, H.W. (1997). The measurement of physical self-concept: A construct validation approach. En K.R. Fox (Eds.), The Psysical Self (pags. 27-58). Human Kinetics.

Marsh, H.W. y Byrne, B.M. (1993). Confirmatory factor analysis of multigroup-multimethod self-concept data: Between-group and within-group invariance constraints. Multivariate Behavioral Research, 28, 313-349.

Marsh, H.W. y Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First and higher order factor models and their invariance across groups. Psychological Bulletin, 97 (3), 562-582.

Marsh, H.W., Balla, J., y McDonald, R.P. (1988). Goodness-of-fit indices in confirmatory factor analysis: The effects of size. Psychological Bulletin, 103, 411-423.

Marsh, H.W., Byrne, B.M., y Shavelson, R. (1988). A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. Journal of Educational Psychology, 80, 366-380.

Marsh, H.W., Hey, J., Roche, L., y Perry, C. (1997). Structure of physical self-concept: Elite athletes and physical education students. Journal of Educational Psychology, 89 (2), 369-380.

Marsh, H.W., Hey, J., y Johnson, S. (en prensa). Elite Athlete Self Description Questionnaire: Hierarchical confirmatory factor analysis of responses by two distinct groups of elite athletes.

Marsh, H.W., Perry, C., Horsely, C., y Roche, L. (1995). Multidimensional self-concepts of elite athletes: How do they differ from the general population? Journal of Sport and Exercise Psychology, 17, 70-83.

Marsh, H.W., y Byrne, B.M. (1993a). Do we see ourselves as others infer: A comparison of self-other agreement on multiple dimensions of self-concept from two continents. Australian Journal of Psychology, 45 (1), 49-58.

Marsh, H.W., y Byrne, B.M. (1993b). Confirmatory factor analysis of multitrait-multimethod self-concept data: Between-group and within-group invariance constraints. Multivariate Behavioral Research, 28, (3), 313-349.

Marsh, H.W., y Hattie, J. (1996). Theoretical perspectives on the structure of self-concept. En B.A. Bracken (Eds.), Handbook of Self-Concept (pags. 38-90). New York: Wiley.

Marsh, H.W., y Jackson, S.A. (1986). Multidimensional self-concepts, maculinity, and feminity as a function of women's involvement in athletics. Sex Roles, 15, 391-415.

Marsh, H.W., y Peart, N.D. (1988). Competitive and cooperative physical fitness training for girls: Effects on physical fitness and multidimensional sef-concepts. Journal of Sports and Exercise Psychology, 10, 390-407.

Marsh, H.W., y Redmayne, R.S. (1994). A multidimensional physical self-concept and its relation to multiples components of physical fitness. Journal of Sport and Exercise Psychology, 16, 45-55.

Marsh, H.W., y Richards, G.E. (1988). The Tennessee Self Concept Scales: Reliability, internal structure, and construct validity. Journal of Personality and Social Psychology, 55, 612-624.

Marsh, H.W., y Shavelson, R.J. (1985). Self-concept: Its multifaceted, hieralchical structure. Educational Psychologist, 20, 107-125.

Marsh, H.W., y Smith, I.D. (1982). Multitrait-multimethod analyses of two self-concept instruments. Journal of Educational Psychology, 74, 430-440.

Marsh, H.W.; Richards, G.E.; Johnson, S.; Roche, L.; y Tremayne, P. (1994). Physical Self-Description Questionnaire: Psychometric properties and a multitrait-multimethod analysis of relations to existing instruments. Journal of Sport and Exercise Psychology, 16, 270-305.

Martínez Arias, R. (1995). Psicometría: Teoría de los tests psicológicos y educativos. Madrid: Síntesis.

Marx, R.W., y Winne, P.H. (1978). Construct interpretations of three self-concept inventories. American Educational Research Journal, 15, 99-108.

Mathes, E.W., y Kahn, A. (1975). Physical attractiveness, happiness, neuroticism, and self-esteem. Journal of Psychology, 90, 27-30.

Matsumoto, D. (1994). People: Psychology from a Cultural Perspective. Pacific Grove, CA: Brooks/Cole.

Mazor, K.M., Clauser, B.E., y Hambleton, R.K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. Educational and Psychological Measurement, 54 (2), 284-291.

McCrae, R.R. y Costa, P.T. (1985). Updating Norman's "adequacy taxonomy": Intelligence and personality dimensions in natural language and in questionnaires. Journal of Personality and Social Psychology, 49, 710-721.

McCrae, R.R. y John, O.P. (1992). An introduction to the five-factor model and its applications. Journal of Personality, 60, 175-215.

McGaw, B. y Jöreskog, K.G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. British Journal of Mathematical and Statistical Psychology, 24, 154-168.

Mead, G.H. (1925). The genesis of the self and social control. International Journal of Ethics, 35, 251-273.

Mead, G.H. (1934). Mind, self, and society. Chicago: University of Chicago Press.

Mellenberg, G.J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-108.

Mellenberg, G.J. (1994). A unidimensional latent trait model for continuous item responses. Multivariate Behavioral Research, 29 (3), 223-236.

Mesquita, B. y Frijda, N.H. (1992). Cultural variations in emotions: a review. Psychological Bulletin, 412, 179-204.

Millsap, R.E. y Everson, H.T. (1991). Confirmatory measurement model comparisons using latent means. Multivariate Behavioral Research, 26, 479-497.

Millsap, R.E. y Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement, 17 (4), 297-334.

Miura, I.T.; Okamoto, Y.; Kim, C.C.; Steere, M.; y Fayol, M. (1993). First graders' cognitive representation of number and understanding of place value: Cross-national comparisons - France, Japan, Korea, Sweden, and the United States. Journal of Educational Psychology, 85 (1), 24-30.

Moghaddam, F.M.; Taylor, D.M.; y Wright, S.C. (1993). Social Psychology in Cross-Cultural Perspective. New York: Freeman.

Moreno, Y. (1997). Propiedades psicométricas del Perfil de Autopercepción Física (PSPP). Tesis doctoral no publicada. Valencia: Universitat de València.

Muñiz, J. (1992). Teoría clásica de los tests. Madrid: Pirámide.

Muthén, B. y Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. Psychometrika, 46, 407-419.

Navas Ara, M.J. (1996). Equiparación de puntuaciones. En J. Muñiz (coor.), Psicometría (p. 293-369). Madrid: Universitas.

Nesselroade, J.R. (1994). Exploratory factor analysis with latent variables and the study of processes of development and change.

En A. von Eye y C.C. Clogg (Eds.), Latent variables analysis (pp. 131-154). Thousand Oaks, CA: Sage.

Newcomb, A.F., y Bukowski, W.M. (1983). Social impact and social preference as determinants of children's peer group status. Developmental Psychology, 19, 856-867.

Oakland, T. y Hu, S. (1992). The top 10 tests used with children and youth worldwide. Bulletin of the International Test Commission, 19 (1), 99-120.

Obeyesekere, G. (1981). Medusa's Hair: An Essay on Personal Symbols and Religious Experience. University Chicago Press. 217 pp.

Offer, D., Ostrov, E,. Howard, K.I., y Dolan, S. (1992). Offer Self-Image Questionnaire, Revised. Los Angeles, CA: Western Psychological Services.

Oort, F.J. (1992). Using restricted factor analysis to detect item bias. Methodika, VI, 150-166.

Oort, F.J. (1996). Using restricted factor analysis in test construction. Tesis doctoral no publicada. Amsterdam: Universidad de Amsterdam.

Osterlind, S.J. (1979). Test item bias. Beverly Hills, CA: Sage Publications.

Pallas, A.M., Entwisle, D.R., Alexander, K.L., y Weinstein, P. (1990). Social structure and the development of self-esteem in young children. Social Psychology Quarterly, 53 (4), 302-315.

Pandey, J.; Sinha, D.; Bhawuk, D.P.S. (eds.) (1995). Asian Contributions to Cross-Cultural Psychology. New Delhi: Sage.

Peterson, M.F.; Smith, P.B.; Akande, D.; Ayestaran, S.; y Bochner, S. (1995). Role stress by national culture and organizational function: a 21 nation study. Academical Management Journal, 38, 429-452.

Piers, E. (1969). Manual for the Piers-Harris Children's Self Concept Scale. Nashville: Counselor Recordings and Test.

Piers, E. (1984). Piers-Harris Children's Self Concept Scale: Revised manual. Los Angeles, CA: Western Psychological Services.

Plake, B.S. (1980). A comparison of a statistical and subjective prodecure to ascertain item validity: One step in the test validation process. Educational and Psychological Measurement, 40, 397-404.

Poortinga, Y. (1992). Toward a conceptualization of culture for psychology. Ver Iwawaki et al. 1992, pp. 3-17.

Poortinga, Y.H. (1971). Cross-cultural comparison of maximum performance tests: Some methodological aspects and some experiments with simple auditory and visual stimuli. Psychologia Africana, Monograph Supplement, N° 6.

Poortinga, Y.H. (1983). Psychometric approaches to intergroup comparison: The problem of equivalence. En S.H. Irvine & J.W. Berry (Eds.), Human assessment and cross-cultural factors (pp. 237-258). New York: Plenum.

Poortinga, Y.H. (1989). Equivalence of cross-cultural data: An overview of basic issues. International Journal of Psychology, 24, 737-756.

Pyle, R.L., Mitchell, J.E., y Lokert, E.D. (1981). Bulimia: A report of 34 cases. Journal of Clinical Psychiatry, 42, 60-64.

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 (3), 552-566.

Rengel, E. (1986). Agreement between statistical and judmental item bias methods. Paper presented at he annual meeting of the American Psychological Associationa, Washington, DC. (ERIC Document Reproduction N°. ED 289 890).

Richards, G.E. (1987). Outdoor education in Australia in relation to the Norman Conquest, a Greek olive grove and the external perspective of a horse's mouth. Trabajo presentado en la Fifth National Outdoor Education Conference, Perth, Western, Australia.

Richards, G.E. (1988). Physical Self-Concept Scale. Sydney: Australian Outward Bound Foundation.

Richards, I. (1953). Toward a theory of translation. Studies in Chinese thought. American Anthropological Association, 55, Memoir 75, Chicago: University of Chicago Press.

Robie, C. y Ryan, A.M. (1996). Structural equivalence of a measure of cross-cultural adjustment. Educational and Psychological Measurement, 56 (3), 514-521.

Rogers, C.R. (1951). Client-centered therapy. New York: Houghton Mifflin.

Rogers, H.J. y Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, 17 (2), 105-116.

Roid, G.H., y Fitts, W.H. (1988). Tennessee Self-Concept Scale, Revised manual. Los Angeles, CA: Western Psychological Services.

Rosansky, E.J. (1979). A review of the bilingual Syntax Measure. En B. Spolsky (ed.), Some major tests: Advances in language testing. Series 1. Arlington, VA: Center for Applied Linguistics.

Rosenberg, M. (1965). Society and the adolescent self-image. Princeton, NJ: Princeton University Press.

Rosenberg, M. (1979). Conceiving the self. New York: Basic Books.

Rubinstein, G. (1996). Two peoples in one land. A validation study of Altemeyer's Right-Wing Authoritarianism scale in the Palestinian and Jewish societies in Israel. Journal of Cross-Cultural Psychology, 27 (2), 216-230.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monographs, 34 (Suppl. 17).

Sandoval, J. y Miille, M.P.W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. Journal of Consulting and Clinical Psychology, 48, 249-253.

Santisteban, C. (1990). Psicometría. Teoría y práctica en la construcción de tests. Madrid: Ediciones Norma.

Sarbin, T.R. (1952). A preface to a psychological analysis of the self. Psychologial Review, 59, 11-22.

Schecter, M.D. (1961). The orthopaedically handicapped child: Emotional reactions. Archives of General Psychiatry, 4, 247-253.

Scheuneman, J.D. (1987). An experimental, exploratory study of causes of bias in test items. Journal of Educational Measurement, 24, 97-118.

Schwartz, S.H. (1992). The universal content and structure of values: theoretical advances and empirical tests in 20 countries. En Advanced Experimental Social Psychology, ed. MP Zanna, 25, 1-65. New York: Academic.

Schwartz, S.H. (1994). Studying human values. En A. Bouvy, F.J.R. Van de Vijver, P. Boski, & P. Schmitz (Eds.), Journeys into cross-cultural psychology (p. 239-254). Lisse: Swets & Zeitlinger.

Segall, M.H.; Campbell, D.T.; y Herskovits, M.J. (1966). The influence of culture on visual perception. Indianapolis, IN: Bobbs-Merrill.

Segall, M.H.; Dasen, P.R.; Berry, J.W.; y Poortinga, Y.H. (1990). Human Behavior in Global Perspective: An Introduction to Cross-Cultural Psychology. New York: Pergamon.

Shavelson, R.J., Hubner, J.J., y Stanton, G.C. (1976). Validation of construct interpretations. Review of Educational Research, 46, 407-441.

Shavelson, R.J., y Bolus, R. (1982). Self-concept: The interplay of theory and methods. Journal of Educational Psychology, 74, 3-17.

Shavelson, R.J., y Marsh, H.W. (1986). On the structure of self-concept. En R. Schwarzer (Ed.), Anxiety and cognitions. Hillsdale, NJ: Erlbaum.

Shaver, P.R.; Wu, S.; y Schwartz, J.C. (1991). Cross-cultural similarities and differences in emotion and its representation: a prototype approach. En Review of Personality and Social Psychology, ed. MS Clark, 13, 175-212. Beverley Hills, CA: Sage.

Shepard, L.A., Camilli, G., y Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

Shepard, LA., Camilli, G. y Williams, D.M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Shostak, M. (1983). Nisa: The life and words of a !Kung woman. New York: Vintage. 402 pp.

Shweder, R.A. (1990). Cultural psychology:What is it? Ver Stigler et al. 1990, pp. 1-43. Ver también Shweder 1991, pp. 73-112.

Shweder, R.A. (1991). Thinking Through Culture: Expeditions in Cultural Psychology. Cambridge, Mass.: Harvard University Press. 404 pp.

Shweder, R.A. y Sullivan, M. (1990). The semiotic subject of cultural psychology. En Handbook of Personality: Theory and Research, ed. L.A. Pervin, pp 399-416. New York: Guilford.

Shweder, R.A. y Sullivan, M.A. (1993). Cultural psychology: Who needs it? Annual Review of Psychology, 44, 497-523.

Simmons, R.G., y Rosenberg, F. (1975). Sex, sex-roles, and self-image. Journal of Youth and Adolescence, 4, 229-258.

Smith, C.S.; Tisak, J.; Bauman, T.; y Green, E. (1991). Psychometric equivalence of a translated circadian rhythm questionnaire: Implications for between and within population assessments. Journal of Applied Psychology, 76 (5), 628-636.

Smith, P.B. y Bond, M.H. (1994). Social Psychology Across Cultures: Analysis and Perspectives. Boston: Allyn & Bacon.

Soares, L.M., y Soares, A.T. (1977). The self-concept: Mini, maxi, multi. Trabajo presentado en la reunión anual de la American Educational Research Association, New York.

Soares, L.M., y Soares, A.T. (1983). Components of students' self-related cognitions. Trabajo presentado en la reunión anual de la American Educational Research Association, Montreal, Quebec, Canada.

Song, I.S., y Hattie, J.A. (1984). Home environment, self-concept, and academic achievement: A causal modeling approach. Journal of Educational Psychology, 76, 1269-1281.

Sonstroem, R.J., Speliotis, E.D., y Fava, J.L. (1992). Perceived physical competence in adults: An examination of the Physical Self-Perception Scale. Journal of Sport and Exercise Psychology, 10, 207-221.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. British Journal of Mathematical and Statistical Psychology, 27, 229-239.

Sörbom, D. (1978). An alternative to the methodology for analysis of covariance. Psychometrika, 43, 381-396.

Sörbom, D. (1982). Structural equation models with structured means. En K.G. Jöreskog y H. Wold (Eds.), Systems under indirect observation: Causality, structure, prediction (pp. 183-195). Amsterdam: North-Holland.

SPSS Inc. (1993). SPSS for Windows. Base System User's Guide (6.0).

Stein, R.J. (1996). Physical self-concept. En B.A. Bracken (Eds.), Handbook of Self-Concept (pags. 374-394). New York: Wiley.

Stein, R.J., Bracken, B.A., Shadish, W., y Haddock, C.K. (1995). The development of the Children's Physical Self-Concept Scale. En proceso de publicación.

Stigler, J.W.; Shweder, R.; Herdt, G., eds. (1990). Cultural Psychology: Essays on Comparative Human Development. Cambridge University Press. 625 pp.

Stocking, M.L. y Lord, F.M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Sullivan, H.S. (1953). The interpersonal theory of psychiatry. New York: Norton.

Super, C.M. (1981). Behavior development in infancy. En R.H. Munroe, R.L. Munroe, y B.B. Whiting (Eds.), Handbook of cross-cultural human development (pp. 181-270). New York: Garland SPTM Press.

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

Thissen, D., Steinberg, L., Pyszczynski, T., y Greenberg J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. Applied Psychological Measurement, 7, 211-226.

Triandis, H.C. (1994). Culture and Social Behavior. New York: McGraw-Hill.

Triandis, H.C.; Dunnette, M.; Hough, L.M. (eds.) (1993). Handbook of Industrial and Organizational Psychology, Vol. 4, Cross-Cultural Studies. Palo Alto, CA: Consulting Psychologists. 2ª ed.

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

Van de Vijver, F. y Leung, K. (1996). Methods and data analysis of comparative research. En J.W. Berry, Y.H. Poortinga y J. Pandey (eds.), Handbook of Cross-Cultural Psychology (2ª ed., Vol. 3, pp. 257-300). Needham, MA: Allyn & Bacon.

Van de Vijver, F.J.R. (1988). Systematizing item content in test design. En R. Langeheine y J. Rost (eds.), Latent trait and latent class models (pp. 291-307). New York: Plenum.

Van de Vijver, F.J.R. (1994). Item bias: Where psychology and methodology meet. En A. Bouvy, F.J.R. Van de Vijver, P. Boski, y P. Schmitz (eds.), Journeys into cross-cultural psychology (pp. 111-126). Lisse: Swets & Zeitlinger.

Van de Vijver, F.J.R. y Lonner, W. (1995). A bibliometric analysis of the Journal of Cross-Cultural Psychology. Journal of Cross-Cultural Psychology, 26, 591-602.

Van de Vijver, F.J.R. y Poortinga, Y.H. (1982). Cross-cultural generalization and universality. Journal of Cross-Cultual Psychology, 13, 387-408.

Van de Vijver, F.J.R. 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: Kluwer.

Van de Vijver, F.J.R. y Poortinga, Y.H. (1992). Testing in culturally heterogeneous populations: When are cultural loadings undesirable? European Journal of Psychological Assessment, 8, 17-24.

Van den Wollenberg, A.L. (1988). Testing a latent trait model. En R. Langeheine y J. Rost (eds.), Latent trait and latent class models (pp. 31-50). New York: Plenum.

Van Haaften, E.H. y Van de Vijver, F.J.R. (in press). Psychological consequences of environmental degradation. Journal of Health Psychology.

Van Leest, P.F. y Bleichrodt, N. (1990). Testing of college graduates from ethnic minority groups. En N. Bleichrodt y P.J. Drenth (Eds.), Comtemporary issues in cross-cultural psychology. Amsterdam, The Netherlands: Swets and Zeitlinger.

Vernon, P.E. (1969). Intelligence and cultural environment. London: Methuen.

Vernon, P.E. (1979). Intelligence: Heredity and environment. San Francisco: Freeman.

Vorst, H.C.M. (1985). Manual of the School Attitude Questionnaire. Nijmegen: Berkhout Nijmegen.

Vorst, H.C.M. (1989). Manual of the School Attitude Questionnaire International. Nijmegen: Berkhout Nijmegen.

Watkins, D. y Dong, Q. (1994). Assessing the self-esteem of Chinese school children. Educational Psychology, 14, 129-137.

Weisner, T.S. (1984). A cross-cultural perspective: ecological niches of middle childhood. En The Elementary School Years: Understanding Development During Middle Childhood, ed. A. Collins, pp. 335-369. Washington: National Academy.

Welch, C.J. y Hoover, H.D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. Applied Measurement in Education, 6 (1), 1-19.

Welch, C.J. y Miller, T.R. (1995). Assessing differential item functioning in direct writing assessments: Problems and an example. Journal of Educational Measurement, 32 (2), 163-178.

Wells, L.E., y Marwell, G. (1976). Self-esteem: Its conceptualization and measurement. Beverly Hills, CA: Sage.

Werner, O. y Campbell, D.T. (1970). Translating, working through interpreters, and the problem of decentering. En R. Naroll y R. Cohen (eds.), A Handbook of Cultual Anthropology (pp. 398-419). New York: American Museum of Natural History.

Williams, J.E. y Best, D.L. (1982). Measuring sex stereotypes: A thirty-nation study. Beverly Hills, CA: Sage.

Williams, J.E. y Best, D.L. (1990). Sex and Psyche: Gender and Self Viewed Cross-Culturally. Newbury Park, CA: Sage.

Wills, W. (1982). The science of translation: Problems and methods. Tuebingen: Narr.

Windle, M., Iwawaki, S., y Lerner, R.M. (1988). Cross-cultural comparability of temperament among Japanese and American preschool children. International Journal of Psychology, 23, 547-567.

Wylie, , R.C. (1989). Measures of self-concept. Lincoln, NE: University of Nebraska Press.

Wylie, R.C. (1974). The self-concept. Rev. ed. Vol. 1. Lincoln, NE: University of Nebraska Press.

Wylie, R.C. (1979). The self-concept. Vol. 2. Lincoln, NE: University of Nebraska Press.

Zwick, R., Donoghue, J.R., y Grima, A. (1993). Assessment of differencial item functioning for performance tasks. Journal of Educational Measurement, 30 (3), 233-251.