Psicothema
2000. Vol. 12, Supl. nº 2, pp. 540-544

ISSN 0214 - 9915 CODEN PSOTEG

Teoría de respuesta al ítem y análisis factorial confirmatorio: dos métodos para analizar la equivalencia psicométrica en la traducción de cuestionarios

Inés Tomás Marco*, Vicente González-Romá* y Juana Gómez Benito**
* Universidad de Valencia y ** Universidad de Barcelona

La traducción de cuestionarios es un fenómeno frecuente en la investigación transcultural. En estos casos, no se puede asumir la equivalencia de las diferentes versiones traducidas, sino que es necesario confirmar dicha equivalencia llevando a cabo análisis del FDI. El análisis factorial confirmatorio (AFC) y la teoría de respuesta al ítem (TRI), representan dos aproximaciones alternativas para el estudio del FDI. El objetivo de este trabajo es comparar el AFC y la TRI como métodos alternativos para evaluar la equivalencia psicométrica en el contexto de la traducción de instrumentos de medida. Para ello se aplicó la versión original en inglés del PSDQ a una muestra de 986 adolescentes australianos, y una versión traducida al castellano de este cuestionario a 986 adolescentes valencianos. Se llevaron a cabo AFCs utilizando el modelo de medias latentes; dentro del contexto de la TRI se utilizó el Modelo de Respuesta Graduada de Samejima. Los resultados indican que el uso del AFC con estructura de medias latentes es totalmente comparable a los métodos de la TRI, y permite analizar tanto el FDI uniforme como el no uniforme.

Item response theory and confirmatory factor analysis: Two approaches for testing the psychometric equivalence of translated tests. Translating psychological tests from one language and culture to other languages and cultures is a common practice in cross-cultural research. However, researchers should not assume that the translation is perfectly equivalent to the original source language version; instead of that it is necessary to carry out DIF analysis to test the equivalence of the different versions of the questionnaire. Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT) models are alternative procedures for the detection of DIF. The objective of this study is to compare the utility of CFA and IRT models for testing the psychometric equivalence of translated tests. The English version of the PSDQ was administered to 986 Australian teenagers, and a translated Spanish version of the questionnaire was administered to 986 Spanish teenagers. CFA with nonzero latent means were carried out to test the equivalence of both versions of the questionnaire; and in the context of IRT Samejima's Graded Response Model was used with the same purpose. The results of this study show that CFA with nonzero latent means allows to identify uniform DIF as well as nonuniform DIF, so the use of AFC is comparable to the use of IRT models.

En la investigación transcultural es frecuente la necesidad de realizar traducciones y adaptaciones de cuestionarios con el fin de realizar comparaciones entre poblaciones de sujetos que no comparten el mismo idioma. En estos casos, es conveniente utilizar los métodos y las recomendaciones que sugiere la literatura transcultural para llevar a cabo la traducción y adaptación de instrumentos de medida (Hambleton, 1994; Hambleton,1996; Hambleton & Kanjee, 1995; Van de Vijver & Hambleton, 1996). Sin embargo, a pesar de seguir estas recomendaciones, no se puede asumir la equivalencia de las diferentes versiones traducidas. Es necesario confirmar dicha equivalencia llevando a cabo análisis que permitan detectar aquellos ítems que presentan funcionamiento diferencial.

El análisis factorial confirmatorio (AFC) y la teoría de respuesta al ítem (TRI), representan dos aproximaciones para abordar el estudio de la equivalencia de versiones en distinto idioma de un mismo cuestionario. Trabajos previos han comparado estos dos procedimientos para el estudio del Funcionamiento Diferencial de los Items (FDI) (Reise, Widaman, & Pugh, 1993; Maurer, Raju, & Collins, 1998). Estos trabajos ofrecen interesantes conclusiones sobre las semejanzas y diferencias tanto a nivel teórico como práctico entre ambos procedimientos, y las ventajas y desventajas de cada uno de ellos. Una de las principales conclusiones que se deriva de estos trabajos, es que ambos procedimientos ofrecen resultados muy similares, con diferencias mínimas en la identificación de los ítems que presentan funcionamiento diferencial. Estas diferencias se atribuyen a que los procedimientos basados en la TRI imponen más restricciones de invarianza en el estudio del FDI. Ambos trabajos han utilizado AFC con estructura de covarianzas,

Correspondencia: Inés Tomás Marco
Facultad de Psicología
Universidad de Valencia
46010 Valencia (Spain)
E-mail: Ines.Tomas@uv.es

que únicamente permite analizar la invarianza de la saturación factorial o parámetro lambda, que representa el parámetro de discriminación del ítem; mientras que la TRI analiza conjuntamente la invarianza del parámetro de discriminación y del parámetro de dificultad del ítem. La incapacidad del AFC con estructura de covarianzas de ofrecer información respecto al parámetro *b*, resulta todavía más limitante si se tiene en cuenta que este procedimiento solamente permite detectar el FDI no uniforme. Sin embargo, esta importante limitación del AFC puede resolverse si se utiliza el modelo de estructura de medias latentes (Millsap & Everson, 1993). Este modelo permite expresar el valor esperado de las variables observables en función de las medias de las variables latentes:

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

donde 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, y 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.

Desde este enfoque, el estudio del FDI implica analizar la invarianza de las matrices de saturaciones factoriales y de los vectores de parámetros interceptos entre los grupos a comparar. Según Mellenbergh (1994), la saturación factorial (λ_x) puede ser interpretada como el parámetro de discriminación del ítem, y el intercepto (τ_x) puede ser interpretado como el parámetro de dificultad del ítem. Se debe tener en cuenta que 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, 1996). 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, & Greenberg (1983), Millsap & Everson (1991), y Everson, Millsap, & Rodriguez (1991), representan algunos ejemplos del uso del AFC con medias latentes utilizando escalas de respuesta tipo Likert.

En el contexto de la TRI, la curva característica de un ítem (CCI) representa la relación entre el valor de respuesta esperado en el ítem y el valor en el rasgo o aptitud medida por el test al que pertenece dicho ítem. Considerando la problemática de la traducción de instrumentos de medida, un ítem presenta funcionamiento diferencial si sus parámetros (y por lo tanto su curva característica) difieren de forma significativa en cada uno de los grupos a los que se han administrado las diferentes versiones del cuestionario. Ello indica que sujetos con el mismo nivel en la aptitud pero que pertenecen a grupos diferentes (y que por lo tanto han completado versiones diferentes del cuestionario), no tienen la misma respuesta esperada en ese ítem. El cuestionario traducido y analizado en este trabajo está compuesto por ítems politómicos, que se responden en una escala con seis categorías de respuesta, graduadas desde «totalmente verdadero» a «totalmente falso», por ello, el modelo utilizado para el análisis de los datos es el Modelo de Respuesta Graduada (MRG) de Samejima (1969).

El objetivo general de este trabajo es comparar el AFC y la TRI como procedimientos alternativos para evaluar la equivalencia psicométrica en el contexto de la traducción de instrumentos de medida. Este objetivo general se desglosa en los siguientes objetivos específicos: a) mostrar la aplicación del AFC con estructura de medias latentes para detectar tanto el FDI no uniforme como el FDI uniforme; y b) comparar los resultados obtenidos con los dos métodos alternativos de detección del FDI: AFC con estructura de medias latentes, y TRI utilizando el modelo de respuesta graduada de Samejima. El punto a representa una aportación novedosa al estudio del FDI en el contexto de la traducción de instrumentos de medida. Su propósito es superar algunas de las limitaciones de trabajos previos respecto a la comparación entre métodos de detección del FDI (el AFC y la TRI).

Metodo

La muestra está compuesta por 986 adolescentes australianos (54.5% varones, 45.5% mujeres), y 986 adolescentes valencianos (50.6% varones, 49.4% mujeres). Respecto a la edad, todos los sujetos tienen edades comprendidas entre los 12 y los 16 años, y son estudiantes de educación secundaria. La media de edad en el grupo australiano es de 13.5 (dt = 1.11), en el grupo español la media de edad es de 13.3, (dt = 1.07).

Como parte de un proyecto más amplio, en este trabajo se analiza la equivalencia psicométrica de la subescala Autoconcepto Físico Global del Physical Self-Description Questionnaire (PSDQ) (Marsh, Richards, Johnson, Roche, & Tremayne, 1994). Esta subescala está compuesta por 6 ítems que se responden en una escala de respuesta tipo Likert que oscila entre 1 (=totalmente falso) y 6 (=totalmente verdadero). En la tabla 1 aparecen los enunciados de los 6 ítems que componen la subescala «Autoconcepto Físico Global» del cuestionario PSDQ en su versión en inglés, y en su versión traducida al castellano (Tomás, 1998). A partir de la versión original en inglés del cuestionario, se llevó a cabo una traducción al castellano, utilizando el procedimiento de traducción inversa o back-translation (Brislin, 1980). En el grupo de adolescentes australianos se administró la versión original en inglés; mientras que en el grupo de adolescentes valencianos se administró la versión traducida al castellano.

Se llevaron a cabo AFCs con el programa LISREL 8 (Jöreskog & Sörbom, 1993) utilizando el modelo de medias latentes. Dada la

Tabla 1
Items de la subescala «Autoconcepto Físico Global» del cuestionario PSDQ (versión en inglés y en castellano)

1. I am satisfied with the kind of person I am physically.
2. Physically, I am happy with myself.
3. I feel good about the way I look and what I can do physically.
4. Physically I feel good about myself.
5. I feel good about who I am and what I can do physically.
6. I feel good about who I am physically.
1. Físicamente, estoy satisfecho/a con el tipo de persona que soy.
2. Físicamente, me siento contento/a conmigo mismo/a.
3. Me siento satisfecho/a con mi apariencia física y con lo que puedo hacer físicamente.
4. Físicamente, me siento satisfecho/a conmigo mismo/a.
5. Me siento satisfecho/a con quien soy y con lo que puedo hacer físicamente.
6. Estoy satisfecho/a con cómo soy físicamente.

naturaleza de la escala de respuesta del cuestionario se utilizaron correlaciones policóricas, aunque al tratarse de análisis multimuestra, la matriz analizada fue la de covarianzas. Dentro del contexto de la TRI se utilizó el modelo de respuesta graduada de Samejima (1969) implementado en el programa MULTILOG 5.1 (Thissen, 1988). Para llevar a cabo la comparación, se pusieron a prueba de forma independiente con los dos procedimientos, los siguientes modelos anidados: Modelo 1 = modelo base (asume que un único rasgo latente subyace a los 6 ítems analizados y los parámetros son estimados libremente); Modelo 2 = invarianza total de los parámetros *a* y de los parámetros *b*; de no confirmarse el modelo anterior, se pondría a prueba un tercer modelo (Modelo 3) = invarianza parcial de los parámetros *a* y de los parámetros *b*.

Resultados

Los resultados obtenidos con los dos procedimientos de detección del FDI utilizados (AFC y TRI) fueron los siguientes. En el AFC se llevaron a cabo análisis multimuestra con el programa LISREL 8. Se partió de un modelo base (Modelo 1) que planteaba la equivalencia de la estructura factorial de las dos versiones (en inglés y en castellano) de la subescala de Autoconcepto Físico Global del PSDQ. Este modelo base asumía la existencia de un único factor latente subyacente a las 6 variables observables o ítems. Las restricciones de equivalencia entre las dos muestras hacían referencia únicamente a la estructura pero no a los valores de los parámetros de los ítems. La saturación factorial del ítem 2 fue fijada a 1.0 en cada uno de los grupos para identificar la escala de la variable latente. Aunque el valor de χ^2 para este modelo (ver tabla 2) fue estadísticamente significativo (χ^2 (18) = 144.98, p < 0.01), los otros índices sugerían que el ajuste del modelo era razonable (RMSEA=0.060; RMSRS=0.015; GFI=0.98; y CFI=0.99). Por lo tanto, el Modelo 1 fue tomado como modelo base para evaluar el ajuste de los otros dos modelos más restrictivos.

En el Modelo 2, todos los parámetros de la matriz lambda-x (parámetros *a*), y todos los parámetros de la matriz tau-x (parámetros *b*), fueron constreñidos a tomar el mismo valor en las dos muestras, es decir, se evaluó la invarianza de todas las saturaciones factoriales y de todos los interceptos a través de los dos grupos. Como este modelo estaba anidado en el Modelo 1, su ajuste fue evaluado en base a la diferencia de los valores obtenidos para la prueba χ^2 en cada uno de los dos modelos. Como puede observarse en la tabla 2, aunque los índices de bondad de ajuste RMSEA (=0.062), RMSRS (=0.033), GFI (=0.98), y CFI (=0.98) fueron satisfactorios, la diferencia entre la χ^2 de este modelo y la del Modelo 1 resultó estadísticamente significativa (χ^2 (11) = 105.5, p < 0.01). Por lo tanto, no se confirmó la hipótesis de la invarianza total de los parámetros *a* y *b*. Al no confirmarse esta hipótesis, se puso a prueba la invarianza parcial de ambos parámetros.

En base a los valores de los índices de modificación (IM) ofrecidos por el programa LISREL 8, se fueron liberando de forma sucesiva aquellos parámetros cuyos IM presentaban el valor más elevado. Cuando se detectaba que uno de los parámetros (*a* ó *b*) de un ítem particular presentaba el mayor IM, se liberaban conjuntamente ambos parámetros. Se repitió sucesivamente esta operación hasta obtener un modelo (Modelo 3) en el que liberando conjuntamente los parámetros λ_x y τ_x de los ítems 6, 5, 3 y 1, la diferencia en los valores de χ^2 respecto al Modelo 1 no resultaba estadísticamente significativa (χ^2 (3) = 4.85, ns). Además, tal como puede verse en la tabla 2, los otros índices de ajuste (RMSEA=0.056; RMSRS=0.015; GFI=0.98; y CFI=0.99) presentaban una ligera mejora respecto a los obtenidos para el Modelo 2, y ofrecían valores comparables a los obtenidos para el Modelo 1. El Modelo 3 de invarianza parcial de los parámetros confirma la hipótesis de la invarianza parcial de los parámetros de los ítems, e indica que únicamente hay 2 ítems (ítems 2 y 4) cuyos parámetros son invariantes en los dos grupos. Los ítems 5, 6, 3, y 1, presentan funcionamiento diferencial, ya que sus parámetros no son invariantes en los dos grupos.

En el MRG de Samejima, uno de los supuestos básicos que debe cumplirse es que los ítems evaluados representen un único rasgo latente. La comprobación de la unidimensionalidad se llevó a cabo realizando análisis de componentes principales con el programa SPSS (1993) en su versión para Windows, de forma separada con los datos de cada grupo cultural. Una vez comprobado el supuesto de la unidimensionalidad de las escalas, en la TRI también se partió de un modelo base (Modelo 1) en el que todos los parámetros eran estimados libremente. En este modelo, la métrica del rasgo o aptitud no se encuentra identificada entre los grupos ya que no se definen ítems de anclaje; sin embargo, el ajuste de este modelo sirve como base para juzgar el ajuste de los sucesivos modelos en los que se imponen constricciones de invarianza a los ítems (Reise, Widaman, & Pugh, 1993). Para mantener cierta congruencia en la comparación entre AFC y TRI, se utilizó el valor del estadísti-

Tabla 2
Parámetros estimados e Indices de Bondad de Ajuste para los modelos analizados utilizando AFC con estructura de medias latentes

		Modelo 1: modelo base		**Modelo 2: invarianza total de los parámetros**		**Modelo 3: invarianza parcial de los parámetros**	
Item	**Parámetro**	**Australia**	**España**	**Australia**	**España**	**Australia**	**España**
Item 1	λ_{11}	0.95	0.92	0.93		0.95	0.92
	τ_{11}	4.48	4.74	4.62		4.54	4.69
Item 2	λ_{21}	1.00	1.00	1.00		1.00	
	τ_{22}	4.58	4.72	4.66		4.66	
Item 3	λ_{31}	1.01	0.98	0.99		1.01	0.98
	τ_{33}	4.18	4.47	4.33		4.24	4.41
Item 4	λ_{41}	1.09	1.09	1.08		1.09	
	τ_{44}	4.42	4.55	4.49		4.49	
Item 5	λ_{51}	1.02	0.90	0.96		1.02	0.90
	τ_{55}	4.55	4.76	4.66		4.61	4.71
Item 6	λ_{61}	1.06	1.10	1.06		1.06	1.10
	τ_{66}	4.50	4.44	4.50		4.57	4.38
χ^2		144.98*		250.48*		149.83*	
gl		18		29		21	
$\Delta\chi^2$ (vs.Modelo 1)		—		105.5*		4.85 ns	
Δgl(vs. Modelo 1)		—		11		3	
RMSEA		0.060		0.062		0.056	
SRMSR		0.015		0.033		0.015	
GFI		0.98		0.98		0.98	
CFI		0.99		0.98		0.99	

* = p < 0.01; ns = no significativo.
Nota: los parámetros estimados que aparecen centrados entre las columnas de Australia y España, representan parámetros constreñidos a ser invariantes en los dos grupos.

co G^2 para evaluar el ajuste de las CCI utilizando el modelo de respuesta graduada de Samejima. Este estadístico representa «2 veces el logaritmo de la función de probabilidad», y es parte del output ofrecido por el programa MULTILOG 5.1. Bajo ciertas condiciones, G^2 se distribuye como una ji-cuadrado con grados de libertad igual al número de patrones de respuesta menos el número de parámetros estimados en el modelo (Reise et al., 1993). El valor del estadístico G^2 del Modelo 1, sirvió para comparar el ajuste de los sucesivos modelos anidados que fueron puestos a prueba (ver tabla 3).

Tabla 3
Parámetros estimados e Indices de Bondad de Ajuste para los modelos analizados utilizando el modelo de respuesta graduada de Samejima

		Modelo 1: modelo base		**Modelo 2: invarianza total de los parámetros**		**Modelo 3: invarianza parcial de los parámetros**	
Item	**Parámetro**	**Australia**	**España**	**Australia**	**España**	**Australia**	**España**
Item 1	a	2.14	2.07	2.09		2.13	2.05
	b_1	-2.23	-2.58	-2.40		-2.29	-2.54
	b_2	-1.61	-1.84	-1.72		-1.66	-1.79
	b_3	-0.84	-1.20	-1.00		-0.89	-1.14
	b_4	-0.18	-0.49	-0.32		-0.22	-0.43
	b_5	0.73	0.44	0.60		0.70	0.50
Item 2	a	2.30	2.82	2.48		2.27	2.80
	b_1	-2.03	-2.33	-2.16		-2.08	-2.28
	b_2	-1.60	-1.74	-1.67		-1.65	-1.68
	b_3	-0.92	-1.13	-1.01		-0.97	-1.07
	b_4	-0.21	-0.47	-0.33		-0.25	-0.41
	b_5	0.59	0.48	0.55		0.55	0.54
Item 3	a	2.42	2.60	2.48		2.41	2.59
	b_1	-1.93	-2.27	-2.09		-1.99	-2.22
	b_2	-1.40	-1.65	-1.52		-1.45	-1.60
	b_3	-0.60	-0.88	-0.73		-0.64	-0.82
	b_4	0.24	-0.16	0.05		0.20	-0.99
	b_5	1.03	0.84	0.96		1.01	0.90
Item 4	a	3.36	3.38	3.73		3.35	
	b_1	-1.82	-2.04	-1.93		-1.93	
	b_2	-1.33	-1.49	-1.41		-1.41	
	b_3	-0.69	-0.88	-0.77		-0.77	
	b_4	-0.09	-0.26	-0.16		-0.16	
	b_5	0.68	0.56	0.63		0.63	
Item 5	a	2.86	2.35	2.61		2.84	2.38
	b_1	-2.00	-2.57	-2.23		-2.05	-2.50
	b_2	-1.57	-1.89	-1.71		-1.63	-1.83
	b_3	-0.86	-1.21	-1.02		-0.91	-1.15
	b_4	-0.19	-0.52	-0.33		-0.23	-0.46
	b_5	0.59	0.53	0.57		0.56	0.58
Item 6	a	3.42	2.96	3.11		3.39	2.96
	b_1	-1.09	-1.94	-1.93		-1.95	-1.88
	b_2	-1.43	-1.43	-1.43		-1.49	-1.37
	b_3	-0.81	-0.78	-0.79		-0.86	-0.72
	b_4	-0.16	-0.20	-0.16		-0.20	-0.13
	b_5	0.66	0.62	0.66		0.63	0.68
G^2		4044.6		4183.8		4056.8	
gl		1098		1134		1104	
ΔG^2 (vs.Modelo 1)		—		139.2*		12.2 ns	
Δgl(vs. Modelo 1)		—		36		6	

* = $p < 0.01$; ns = no significativo.
Nota: los parámetros estimados que aparecen centrados entre las columnas de Australia y España, representan parámetros constreñidos a ser invariantes en los dos grupos.

Para poder comparar los resultados obtenidos aplicando un modelo de TRI con los obtenidos con AFC, en el Modelo 2 se evaluó de nuevo la invarianza total de los parámetros. Como este modelo se encontraba anidado en el Modelo 1, su ajuste pudo ser evaluado en base a la diferencia de los valores de la prueba G^2 en los dos modelos. Como puede observarse en la tabla 3, el incremento de G^2 entre el modelo base y el modelo de invarianza total de los parámetros resultó estadísticamente significativo (G^2 (36) = 139.2, $p < 0.01$). Este resultado indicaba la no confirmación de la hipótesis de la invarianza total de los parámetros *a* y *b*, o lo que es lo mismo, la existencia de algún/os ítem/s que presentaba/n FDI.

En AFC, la identificación de los ítems que presentan FDI se ve facilitada por los índices de modificación (IM) ofrecidos por el programa LISREL 8. Sin embargo, en la TRI la identificación de estos ítems es más tediosa y compleja, ya que no se dispone de ningún índice que indique qué parámetro se debería liberar para mejorar el ajuste del modelo. En el Modelo 3 de invarianza parcial de los parámetros se siguió el mismo procedimiento utilizado en otros trabajos previos (Maurer et al., 1998; Reise et al., 1993): se analizaba ítem a ítem cuál sería la mejora del ajuste del modelo si se liberaban conjuntamente el parámetro *a* y los cinco parámetros *b* de cada uno de los ítems. Cuando se detectaba que la liberación de los parámetros (*a* y *b*) de un ítem particular ofrecía una mejora sustancial en el ajuste del modelo, se liberaban ambos parámetros. Se repitió sucesivamente esta operación hasta obtener un modelo en el que liberando los parámetros de los ítems 6, 2, 3, 5 y 1, la diferencia en los valore de G^2 respecto al Modelo 1 no resultaba estadísticamente significativa (G^2 (6) = 12.2, ns). El Modelo 3 de invarianza parcial de los parámetros confirma la hipótesis de que no todos los parámetros de los ítems son invariantes entre los dos grupos, y tal como se puede ver en la tabla 3, indica que únicamente hay un ítem (ítem 4) cuyos parámetros son invariantes. Los ítems 6, 2, 3, 5, y 1, presentan funcionamiento diferencial, ya que sus parámetros no son invariantes en los dos grupos.

Conclusiones

El objetivo de este trabajo era comparar el AFC y la TRI como procedimientos alternativos para evaluar la equivalencia psicométrica en el contexto de la traducción de instrumentos de medida. El uso de ambos procedimientos en la detección del FDI ha sido comparado tanto a nivel teórico como a nivel práctico en trabajos previos (Maurer et al., 1998; Reise et al., 1993). En estos trabajos se criticaba que el AFC imponía menos exigencias en la evaluación de la equivalencia ya que ignoraba el parámetro *b*. Sin embargo, en este estudio se ha puesto de manifiesto que el empleo del AFC con estructura de medias latentes permite analizar también la invarianza de los parámetros *b*, por lo que es comparable a los métodos de la TRI en lo que se refiere a las restricciones impuestas a los parámetros. De este modo, se ha mostrado cómo el AFC permite analizar tanto el FDI uniforme como el no uniforme.

Por otra parte, y al igual que en trabajos previos (Maurer et al., 1998; Reise et al., 1993), se puede concluir que los resultados obtenidos con ambos métodos son muy similares. Tanto los resultados del AFC como los resultados de la TRI indican que los ítems 1, 3, 5 y 6 presentan FDI, mientras que el ítem 4 no presenta funcionamiento diferencial. La discrepancia aparece respecto al ítem 2, ya que los resultados obtenidos utilizando el MRG de Samejima indican que el ítem 2 presenta FDI, mientras que utilizando AFC los resultados indican que los parámetros de este ítem son in-

variantes en los dos grupos. En trabajos previos, las diferencias observadas en los resultados obtenidos con estos dos métodos se achacaban a que el AFC al ignorar el parámetro *b* era menos exigente en las restricciones impuestas. Sin embargo, con este trabajo se ofrece evidencia empírica de que presentando ambos procedimientos el mismo nivel de restricciones impuestas a los parámetros, la TRI parece ser más exigente en la evaluación de la equivalencia.

En base a los resultados obtenidos en este estudio, se puede concluir que tanto el AFC con estructura de medias latentes como la TRI, ofrecen resultados similares en el análisis del FDI dentro del contexto de la traducción de instrumentos de medida, aunque la TRI parece ser más exigente en la evaluación de la equivalencia. Sin embargo, y ya que ambos procedimientos ofrecen la misma información, por razones prácticas parece ser más aconsejable el uso del AFC. Somos conscientes de que el MRG de Samejima, ha sido desarrollado especificamente para ser utilizado con escalas de respuesta graduada; mientras que el uso del AFC con estructura de medias latentes ha sido propuesto en el contexto de un modelo de respuesta continua al ítem (Mellenbergh, 1994; Ferrando, 1996). No obstante, como se ha indicado anteriormente, el uso del AFC con estructura de medias latentes se puede extrapolar a escalas de respuesta graduada como la escala tipo Likert si se considera que la escala de respuesta continua representa un caso extremo de la escala de respuesta graduada con un elevado número de opciones de respuesta (Mellenberg, 1994).

Referencias

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.

Everson, H.T., Millsap, R.E., & 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. (1996). Calibration of invariant item parameters in a continuous item response model using the extended Lisrel measurement submodel. *Multivariate Behavioral Research*, 31 (4), 419-439.

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. & 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.

Jöreskog, K.G. & Sörbom, D. (1993). *LISREL VIII: User's reference guide*. Mooresville, IN: Scientific Software.

Marsh, H.W., Richards, G.E., Johnson, S., Roche, L., & 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.

Maurer, T.J., Raju, N.S., & Collins, W.C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology*, 83 (5), 693-702.

Mellenberg, G.J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, 29 (3), 223-236.

Millsap, R.E. & Everson, H.T. (1991). Confirmatory measurement model comparisons using latent means. *Multivariate Behavioral Research*, 26, 479-497.

Millsap, R.E. & Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 17 (4), 297-334.

Reise, S.P., Widaman, K.F., & Pugh, R.H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, 114 (3), 552-566.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs*, 34 (Suppl. 17).

SPSS Inc. (1993). *SPSS for Windows*. Base System User's Guide (6.0).

Thissen, D. (1988). *MULTILOG: Multiple, categorical item analysis and test scoring using item response theory* (Version 5.1). Chicago: Scientific Software.

Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement*, 7, 211-226.

Tomás, I. (1998). Equivalencia psicométrica de una traducción del cuestionario de autoconcepto físico PSDQ (Physical Self-Description Questionnaire) al castellano. Tesis doctoral no publicada. Valencia: Universitat de València.

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