Psicológica (1999) 20, 23-40.

Funcionamiento diferencial de los ítems y sesgo en la adaptación de dos pruebas verbales[*]

Paula Elosúa y Alicia López

Universidad del País Vasco

En este trabajo se subrayan las fases a las que debería someterse todo proceso de adaptación de pruebas psicológicas y se ilustran con el estudio de la adaptación al euskera de dos pruebas con componentes verbales. En la primera fase, de carácter exploratorio, se comparan los coeficientes de fiabilidad y las estructuras factoriales y se detecta el funcionamiento diferencial de los ítems. En la segunda fase de índole confirmatoria, se analizan las causas de este último concluyendo la existencia o inexistencia de sesgo. El funcionamiento diferencial se evalúa con el estadístico Mantel-Haenszel y tres medidas derivadas de la teoría de respuesta a los ítems (χ^2 de Lord y las áreas exactas con signo y sin signo de Raju). El alto porcentaje de funcionamiento diferencial encontrado lleva a la conclusión de que las adaptaciones, lejos de ser simples traducciones lingüísticas, han de tener en cuenta distintas dimensiones semánticas que garanticen la equivalencia en el grado de familiaridad y significatividad entre los términos utilizados.

Palabras clave funcionamiento diferencial del ítem, sesgo, equivalencia métrica, adaptación, teoría de respuesta al ítem, área exacta con signo, área exacta sin signo, Mantel-Haenszel, χ^2 de Lord.

La generalización de constructos psicológicos conlleva la necesidad de estandarizar los instrumentos utilizados para su medición más allá de culturas o idiomas específicos. Esta exigencia se traduce básicamente en la adaptación de pruebas psicológicas que garanticen a través de la evaluación de su **equivalencia métrica,** la igualdad de significado entre puntuaciones obtenidas por instrumentos originales y traducidos. El análisis de la equivalencia métrica trasciende el mero estudio de idoneidad de la traducción

[*] Este trabajo ha sido financiado por Universidad del País Vasco. UPV 109.231-HA093/96. Correspondencia dirigirla a Paula Elosúa. Facultad de Psicología. Universidad del País Vasco. Avda. Tolosa, 70. 20009 San Sebastián. TELF. 943-448000 X 5693. E-MAIL:pspelolp@sc.ehu.es

lingüística que por sí misma no asegura ni garantiza la equivalencia psicométrica de las pruebas y no puede reducirse a la comparación de coeficientes de fiabilidad, coeficientes de validez o estructuras factoriales. Pues si bien su igualdad es necesaria para la existencia de equivalencia métrica, no es sin embargo una condición suficiente. Al análisis comparativo entre instrumentos de medida como conjunto o bloque cerrado de ítems, es necesario añadir un estudio pormenorizado de sus componentes individuales que evalúe el funcionamiento diferencial como paso previo para la posterior evaluación del sesgo o falta de equivalencia métrica.

El objetivo de toda traducción, sea literaria, pragmática, estetico-poética o etnográfica (Casagrande, 1954) es reproducir en el idioma terminal el objeto de traducción que está dado en el lenguaje original o fuente. Es un proceso de decodificación-recodificación de un mensaje (ver figura 1), que ha de cumplir en los dos idiomas los mismos objetivos. En la fase de decodificación es necesario aprehender el contenido y el sentido del texto, para posteriormente, con la ayuda de las normas lingüísticas, registros y convenciones del lenguaje terminal, lograr una recodificación correcta y adecuada con un estricto control sobre la tipografía, ortografía, morfología, léxico, corrección gramatical, adecuación y coherencia.

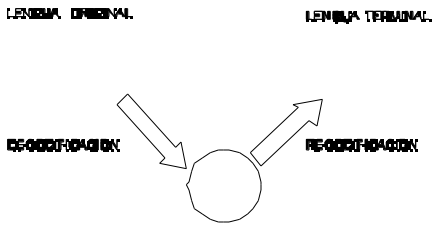

Figura 1. Proceso de traducción.

El estudio de este proceso en el campo de la psicología exige el análisis de la equivalencia métrica entre dos instrumentos de medida (original y adaptado). La equivalencia se logra cuando la relación entre las puntuaciones observadas y la variable latente medida por el test es idéntica entre poblaciones (Drasgow, 1984), o cuando los ítems son psicométricamente equivalentes, es decir, cuando evocan entre un conjunto de respuestas posibles, la misma respuesta específica con igual probabilidad entre sujetos con el mismo nivel de aptitud (Hulin, 1987).

De estas definiciones se deriva la obligatoriedad en todo proceso de adaptación, de efectuar un análisis pormenorizado de las relaciones entre cada uno de los ítems y el rasgo o habilidad que miden. Esta relación se estudia de manera explícita con los modelos de Teoría de Respuesta a los Items (TRI) (Lord, 1980), que proporcionan el marco teórico adecuado para su análisis a

través del estudio de las curvas características de los ítems (CCI) o funciones de respuesta. La curva característica del ítem es una función matemática que establece la relación existente entre las puntuaciones de las personas en la variable medida (θ) y la probabilidad de responder correctamente al ítem ($P_i(\theta)$).

La aportación fundamental de los modelos de TRI en el campo de las adaptaciones de instrumentos de medida psicopedagógicos se deriva de su adecuación para el estudio del funcionamiento diferencial de los ítems (FDI). Un ítem presenta FDI cuando la probabilidad de ser resuelto correctamente por individuos con el mismo nivel en el rasgo varía en función de su grupo de pertenencia (sexo, cultura, nivel socioeconómico..), o lo que es lo mismo cuando sus funciones de respuesta o curvas características de los ítems son diferentes para distintas poblaciones. La correspondencia entre esta definición y las citadas por Drasgow y Hulin, nos lleva a considerar que el estudio de la equivalencia psicométrica de los ítems comienza con la evaluación de su posible funcionamiento diferencial. La TRI ofrece distintos métodos para la detección del FDI que básicamente comparan las respuestas dadas a un ítem por sujetos a los que se les estima un mismo nivel de habilidad en dos grupos distintos, definidos como grupo de referencia y grupo focal (Holland y Thayer, 1988). Entre los procedimientos más utilizados podríamos mencionar, la comparación de parámetros de las curvas características de los ítems en las poblaciones de referencia y focal, (Lord, 1980), el análisis del área limitada por dos curvas características estimadas en dos poblaciones (Rudner, Getson y Knight, 1980; Raju, 1988, 1990), la comparación de curvas empíricas (Hulin, Drasgow y Komocar, 1982) o el estudio de los residuales estandarizados (Linn y Harnisch, 1981).

A pesar de que en la actualidad se dispone de una amplia tecnología para el análisis del funcionamiento diferencial del ítem, y aunque desde un punto de vista teórico son claras las ventajas que aporta su aplicación en el proceso de adaptación de pruebas, hemos de decir que por el momento son escasos los estudios empíricos en los que se hace uso de ella (Bontempo, 1993; Budgell, Raju y Quartetti, 1995;Candell y Hulin, 1986; Drasgow y Hulin, 1989; Drasgow y Lissak, 1983; Ellis, 1989, 1991; Ellis, Becker y Kimmel, 1993; Ellis, Minsel y Becker, 1989; Elosua, López y Torres, 1999; Hambleton y Bollwark, 1991; Hulin, 1987; Hulin, Drasgow y Komocar, 1982; Hulin, Drasgow y Parson, 1983; Hulin y Mayer, 1986). Enmarcado dentro de esta nueva línea de investigación, el objetivo de este trabajo es analizar cada una de las fases por las que ha de pasar toda adaptación de pruebas psicológicas, enfatizando las ventajas que aporta el concepto de funcionamiento diferencial de los ítems.

En todo proceso de adaptación diferenciamos dos etapas; una fase exploratoria en la que se analizan la fiabilidad y validez de las pruebas,

incluyendo dentro del estudio de validez la detección del funcionamiento diferencial de los ítems, y una segunda, a la que llamamos confirmatoria, en la que se buscan explicaciones plausibles que justifiquen las divergencias encontradas entre las versiones fuente y adaptada. En esta segunda parte del proceso se recabaría información sobre el origen de las diferencias para: 1.-analizar el modo de superarlas revisando el contenido de los ítems y corrigiendo aquéllos que presentan problemas. 2.- concluir la existencia de sesgo y por tanto falta de validez de constructo de la prueba adaptada. 3- explicar los resultados como un problema derivado de todo proceso de inferencia basado en la utilización de técnicas estadísticas.

Para ilustrar este proceso se evalúa la adaptación al euskera de dos pruebas con componentes verbales en las que se siguen las pautas de detección del FDI (exploración del sesgo) y búsqueda de las causas del mismo (confirmación del sesgo). Además se evalúa la efectividad y concordancia de distintos procedimientos de detección del FDI. Tres derivan directamente de los modelos de teoría de respuesta a los ítems, área exacta con signo y área exacta sin signo de Raju (1988, 1990), χ^2 de Lord (1980) y el cuarto se basa en el estudio de tablas de contingencia, el estadístico Mantel-Haenszel (Holland y Thayer, 1988).

METODO

Participantes. Los datos se han obtenido de una muestra formada por 1480 sujetos que cursan 4º, 5º y 6º de enseñanza primaria, con edades comprendidas entre los 9 y 11 años. Los sujetos están repartidos por todo el territorio de la Comunidad Autónoma Vasca con el fin de reflejar la heterogeneidad lingüística de cada uno de los territorios históricos.

Del total de la muestra 935 corresponden al grupo bilingües euskaldunes (muestra D). Los 545 restantes forman el grupo de monolingües castellanos (muestra A). La denominación y selección de las muestras se ha determinado por los modelos lingüísticos definidos por la ley del 24 de Diciembre 10/1982.

- **Modelo A**. Todas las asignaturas salvo el euskera se impartirán básicamente en castellano. El euskera tendrá el tratamiento de cualquier otra asignatura.
- **Modelo D**. Todas las asignaturas, salvo el castellano, se darán principalmente en euskera. El euskera también se impartirá como asignatura.

El diseño empleado en la recogida de datos es el que Hambleton (1993) define como "sujetos monolingües en castellano y euskera realizan la prueba original y adaptada". El grupo de referencia lo forman los sujetos

cuya lengua materna es el castellano y están siendo educados en el modelo A. El grupo focal lo forman los sujetos cuya lengua materna es el euskera y están siendo educados en el modelo D.

Instrumentos. La selección de los instrumentos analizados no se ha hecho en función de su calidad psicométrica, sino en el interés expresado por un colectivo de psicólogos escolares y psicopedagogos que hacen uso de las pruebas que vamos a analizar.

Los instrumentos estudiados son las pruebas de aptitud numérica y comprensión verbal integradas en la batería de aptitudes general y aplicada (BADYG) (Yuste, 1988) en su versión elemental (E).

- Aptitud numérica: Es un test de 25 ítems con 5 alternativas de respuesta de las que sólo una es correcta. El objetivo de la prueba es medir el razonamiento numérico, la aplicación de operaciones numéricas en problemas lógico-numéricos y la maduración de funciones matemáticas básicas.

- Comprensión verbal: Consta de 30 ítems y 5 alternativas de respuesta cada uno, con sólo una opción correcta. Los ítems se clasifican según el autor en sinónimos (6), antónimos (5), analogías verbales (12), definición más exacta (5) y finalidad y uso más común (2).

Los valores originales de los coeficientes de fiabilidad calculados por el autor (Yuste, 1988) en las muestras originales son de 0,86 para la prueba de aptitud numérica y de 0,84 para la prueba de comprensión verbal. Ambos coeficientes se calculan con el procedimientos de dos mitades. Cabe mencionar que en el manual técnico de la prueba no se citan los valores de alpha.

Adaptación de la prueba. El proceso de adaptación ha seguido las pautas definidas por Brislin (1970):

- La prueba original escrita en castellano se traduce al euskera por un grupo de licenciados bilingües. Es una traducción fundamentalmente literal.

-Un grupo de licenciados bilingües ajeno al anterior retrotraduce la prueba al castellano.

- Se analizan las diferencias entre las versiones original y retrotraducida con la colaboración de un profesor de enseñanza primaria, de modo que sea la idoneidad el criterio que guíe la solución de las divergencias que pudieran aparecer.

- Un traductor profesional analiza la exactitud y coherencia lingüística de la versión adaptada corrigiendo errores y deficiencias.

RESULTADOS

Los primeros análisis tienen por finalidad describir las muestras y estudiar la consistencia interna de cada una de las pruebas. Los resultados se muestran resumidos en la tabla 1.

Tabla 1. Descripción de las muestras y consistencia interna de las pruebas

	n		N	$\overline{X}$	S_x	α
Aptitud Numérica	25	Modelo A	542	17,22	4,30	0,80
		Modelo D	935	14,53	4,35	0,79
Comprensión Verbal	30	Modelo A	545	20,67	4,77	0,78
		Modelo D	933	12,30	4,57	0,73

Con esta simple descripción de datos, se aprecia que el rendimiento en las dos pruebas es superior en la muestra modelo A que en la muestra modelo D, siendo además esta diferencia significativa con una F de 1113,21 para la prueba de comprensión verbal y de 132,11 para Aptitud numérica.

La consistencia interna de las pruebas se calcula con el alpha de Cronbach (1951), y la igualdad o desigualdad en las distintas muestras se evalúa con el estadístico de Feldt (1969). Los valores obtenidos son de 0,947 para la prueba de aptitud numérica y de 0,814 para comprensión verbal, por lo que podemos afirmar con un nivel de confianza del 99% que los coeficientes de fiabilidad son equivalentes entre las muestras.

Unidimensionalidad

La aplicación de los modelos de teoría de respuesta a los ítems unidimensionales asumen en su formulación la existencia de un rasgo unidimensional que da cuenta de la ejecución de los sujetos en un conjunto de ítems (Hambleton y Swaminathan, 1985). Entre los distintos procedimientos disponibles para evaluar la condición de unidimensionalidad optamos por aquellos que se derivan del análisis de componentes principales.

Ante el problema de la sobreestimación de los factores que deriva de la utilización de variables dicotómicas, y teniendo en cuenta por un lado, que recientes trabajos de simulación ponen de manifiesto que el empleo de correlaciones tetracóricas frente a las correlaciones phi no soluciona totalmente el problema cuando los ítems difieren en dificultad, y por otro, que el objetivo del análisis es únicamente determinar el número de factores (Ferrando, 1996; López Pina, 1995), optamos por el uso de correlaciones phi.

Se someten a un análisis de componentes principales con rotación varimax las matrices de correlaciones phi entre los elementos que componen cada una de las pruebas. Los resultados se recogen en la tabla 2.

Tabla 2. Estructura factorial

		Modelo A		Modelo D	
	factores	Valores propios	Varianza explicada	Valores propios	Varianza explicada
Aptitud numérica	1	4,55	18,2	4,33	17,3
	2	1,91	7,7	2,27	9,1
	3	1,36	5,5	1,31	5,3
Comprensión verbal	1	4,32	14,4	3,67	12,3
	2	1,53	5,1	1,69	5
	3	1,28	4,3	1,25	4,2

Comprensión verbal: se extraen 10 valores propios mayores que la unidad en la muestra A que explican un 50,3% de la varianza total. En la muestra D son 11 los valores propios mayores que la unidad que dan cuenta del 51,4% de la varianza. En ninguna de las dos muestras se alcanza el criterio de unidimensionalidad de Reckase (1979), según el cual el primer factor ha de explicar el 20% de la varianza total. Si aplicamos uno de los índices de unidimensionalidad propuesto por Lord (1980), la razón entre la diferencia de los dos primeros valores propios y la diferencia entre el segundo y el tercero, se obtienen los índices de 11,16 y 4,5, que no nos hacen más que reforzar la idea de mayor acercamiento a la unidimensionalidad de la muestra Modelo A frente a la muestra modelo D.

Aptitud numérica: Se extraen seis factores con valores propios mayores que la unidad en la muestra A que explican el 45,2% de la varianza. En la muestra D seis factores explican el 45% de la varianza. Tampoco en esta prueba se supera el criterio de Reckase, y los índices propuestos por Lord alcanzan los valores de 4,8 y 2,14 para las muestras A y D respectivamente.

Estimación de los parámetros de los ítems

La estimación de los parámetros y el ajuste de los modelos se efectúa de modo independiente en cada una de las muestras con el programa BILOG3, (Mislevy y Bock, 1990) utilizando el procedimiento de estimación marginal de máxima verosimilitud y evaluando el ajuste de cada uno de los ítems con la prueba estadística de χ^2.

Aptitud numérica. En la muestra Modelo A todos los ítems se ajustan perfectamente al modelo de dos parámetros, ofreciendo la prueba un índice de ajuste general de 171,9 ($p \leq 0,3608$). En la muestra Modelo D los ítems 10, 15, 20, 22 y 25 tienen valores χ^2 con un nivel de significación menor de 0,01 ($p \leq 0,01$), presentado la prueba en su totalidad una $\chi^2 = 347,1$ lo que indica falta de ajuste entre el modelo y los datos.

Se evalúa también el ajuste que proporcionaría la utilización de un modelo logístico de tres parámetros. En la muestra monolingüe castellana el valor de ajuste total es de 189,7, con una probabilidad asociada de 0,0827. En la muestra modelo D sin embargo, son seis los ítems que muestran valores significativos (10,12,13,15 y 18), y que impiden un buen ajuste entre el modelo y los datos.

Comprensión verbal: Si evaluamos el ajuste del modelo logístico de dos parámetros en el modelo A ninguno de los ítems presenta valores χ^2 significativos mientras que en el Modelo D los ítems 13, 16 y 22 obtienen índices significativos ($p \leq 0,01$).

Con respecto al modelo logístico de tres parámetros hemos de decir que la probabilidad asociada al valor de χ^2 de 257,1 es de 0,0208 en la muestra modelo A. En la muestra modelo D sin embargo el valor de ajuste global es de 64838 con una probabilidad asociada de $p \leq 0,000001$.

La falta de mejora en el ajuste del modelo logístico de tres parámetros frente al modelo logístico de dos parámetros y el principio de parsimonia nos llevan a seleccionar este último evitando así

los problemas inherentes a la estimación del parámetro de pseudo-azar (c) (Muñiz, 1990; Kolen, 1981; Thissen y Wainer, 1982).

Equiparación de métricas

En los modelos de TRI la escala θ no es única y cualquier transformación de la misma, no varía las características del modelo. Como consecuencia de esta indeterminación es condición previa y necesaria a la evaluación del funcionamiento diferencial, la equiparación de las escalas de los grupos de referencia y focal. Para ello utilizamos el procedimiento de la curva característica (Stocking y Lord, 1983) implementado en EQUATE2 (Baker, 1993),que ejecutamos anclando las pruebas con los ítems que no presentan problemas de ajuste en ninguna de las muestras. Una vez sometidos los parámetros de los ítems a la misma escala es posible calcular los índices de funcionamiento diferencial.

Funcionamiento diferencial de los ítems

Area exacta de Raju: Para el estudio del funcionamiento diferencial de los ítems se utiliza como índice el área limitada por las curvas características de los ítem calculadas en las poblaciones de referencia y focal. En el caso de que las curvas se superpongan el área será 0 y concluiremos ausencia de funcionamiento diferencial. A medida que el valor calculado se aleje de 0 aumentará el índice de FD. El cálculo de este índice se efectúa con las formulas dadas por el autor (Raju, 1988) área exacta con signo (AECS) y área exacta sin signo (AESS) y su significatividad se contrasta con el estadístico *z* (Raju, 1990).

χ^2 de Lord: Este autor propone un estadístico para contrastar la hipótesis nula de igualdad de los vectores que definen los parámetros de los ítems en las poblaciones de referencia y focal (Lord, 1980). En el caso de que los parámetros de los ítems sean iguales, salvo errores aleatorios, las curvas características derivadas de ellos serán idénticas concluyendo ausencia de funcionamiento diferencial.

Para la detección del funcionamiento diferencial de los ítems se ha llevado a cabo un procedimiento iterativo de purificación de la puntuación (Candell y Drasggow, 1988; Kim y Cohen, 1992a; Park y Lautenschalager, 1990) basado en los trabajos de Candell y Drasgow. Consiste básicamente en estimar los parámetros de los ítems de cada grupo, equiparar las métricas, estimar el funcionamiento diferencial de los ítems, reequiparar las métricas utilizando como test de anclaje el compuesto por los ítems sin funcionamiento diferencial y reestimar el funcionamiento diferencial. Este proceso iterativo concluye cuando en dos fases consecutivas se obtienen los mismos resultados.

El cálculo de estos índices se ha efectuado con la ayuda del programa IRTDIF (Kim y Cohen, 1992b)

Mantel-Haenszel: El estadístico Mantel-Haenszel (MH) (Mantel Haenszel, 1959) es un procedimiento simple para el estudio de tablas de contingencia, que por su parsimonia y eficacia se ha convertido en uno de los procedimientos de detección de funcionamiento diferencial más utilizado. Compara las respuestas dadas a un ítem por sujetos que perteneciendo a distintas poblaciones muestran el mismo nivel de puntuación en el test. La hipótesis nula a contrastar afirmaría la existencia de igualdad entre las proporciones de sujetos que aciertan y fallan el ítem en cada una de las muestras y para cada uno de los niveles en que se ha dividido la puntuación total. Este estadístico sigue una distribución χ^2 con un grado de libertad.

En la detección del funcionamiento diferencial de los ítems se aplica el estadístico MH (Fidalgo, 1994) en dos fases. En la primera de ellas, se calcula la puntuación total con todos los ítems que componen la prueba y se evalúa el índice de funcionamiento diferencial. En una segunda etapa, se

recalcula la puntuación total únicamente con los ítems que carecen de funcionamiento diferencial. Con el criterio interno purificado se vuelven a calcular los índices de FDI.

La tabla 3 resume los datos obtenidos tras la aplicación de estos procedimientos. El asterisco al lado de los ítems refleja aquéllos que presentan funcionamiento diferencial con un nivel de riesgo de 0,05.

Tabla 3. Funcionamiento diferencial de los ítems

Ítem	Aptitud numérica				Comprensión verbal			
	$L\chi^2$	AECS	AESS	χ^2	$L\chi^2$	AECS	AESS	χ^2
1	0.9225	1.6045	2.0099	0.61	1.9320	2.7479	2.9058	7.15*
2	1.5822	1.5346	1.9557	0.02	26.6707*	1.6537*	1.6661*	68.62*
3	0.2399	-0.3659	0.7012	0.26	6.3520*	0.7435	0.7910*	4.66*
4	1.8928	-0.5649	0.6749	2.65	22.3809*	-1.1635*	1.2682	11.80*
5	11.5226*	0.1473	0.1747	6.95*	4.7922	-0.4079	0.4079	0.98
6	2.9277	-0.7379	1.2135	3.11	20.8427*	-0.9001*	1.6799*	0.76
7	1.5012	0.0354	0.2745	0.59	22.7396*	-1.3837*	1.5184	41.73*
8	4.8653	-0.3459	0.3459	3.28	69.0118*	1.0539	2.8180*	59.62*
9	1.7636	-0.1691	0.1753	0.39	22.3582*	0.1558	2.5623*	19.41*
10	2.5370	-0.5634	0.9545	1.57	54.2890*	-0.9694	2.0143*	39.72*
11	48.0551*	1.0632*	1.0635*	88.64*	12.6072*	1.1994*	4.0210*	2.95
12	1.6010	-0.0807	0.3128	4.64*	10.1461*	0.1064	0.9451*	26.39*
13	7.0214*	-0.0613	0.9039	3.24	7.9514*	-0.5336	0.5336	2.64
14	0.0713	0.0318	0.0433	0.42	51.8765*	-0.8993*	0.9038*	10.65*
15	14.1795*	-0.3863*	0.4366*	1.63	12.9130*	3.9887*	3.9976*	122.59*
16	0.9463	-0.1022	0.1030	0.10	0.6103	-0.1392	0.3789	0.23
17	7.8343*	0.3269*	0.3271*	17.86*	11.1997*	1.3717	2.3966*	8.70*
18	8.7928*	-0.2924*	0.3035*	1.37	53.9802*	-0.9235*	1.0114*	17.09*
19	7.1055*	0.2495*	0.2857*	21.18*	106.6731*	-1.3900*	1.9427*	40.42*
20	14.1735*	-0.3645*	0.3645*	2.87	10.2715*	7.2057*	7.5146*	211.74*
21	22.1483*	-0.3793*	0.3793*	1.53	15.6493*	-0.5672*	0.8074*	5.68*
22	3.8383	0.1612	0.1703	19.99*	63.7516*	-0.9747*	1.2058*	26.26*
23	2.6507	-0.2623	0.3606	0.01	76.0514*	-1.3647*	1.6930*	21.54*
24	3.0662	-0.1141	0.3625	0.00	0.6016	-0.1316	0.1334	1.74
25	3.9621	0.0519	0.5685	0.03	1.6547	0.2541	0.2574	8.00*
26					9.2432*	-0.4615*	2.1944*	1.32
27					4.5148	0.7950*	0.9226	26.04*
28					1.4019	0.1127	0.2435	18.72*
29					185.3907*	-1.9744*	2.0577*	76.39*
30					136.8171*	-1.8799*	2.1435*	52.73*
total	9	7	7	6	23	17	20	23

Constantes de equiparación:
Fase 1: A=0.9607
K= -0.7026
Fase 2: A=0.9528
K= -0.6304

Constantes de equiparación:
Fase 1: A=0.8025
K= -1.9304
Fase 2: A=0.8341
K= -2.0798

En la prueba de aptitud numérica, el número de ítems que presentan funcionamiento diferencial varía entre 6 y 9 en función del procedimiento utilizado. Según el estadístico MH el porcentaje de ítems con FD es del 24%

y según la χ^2 de Lord el 36%, lo que equivale a 6 ítems en el primer caso y 9 ítems en el segundo. Las áreas de Raju , con una concordancia total en este caso, detectan 7 ítems. En la prueba de comprensión verbal estos niveles se elevan sustancialmente. El procedimiento que menos detecciones muestra es el área exacta de Raju con 17 valores significativos mientras que el procedimiento de Lord y el MH detectan un 76% de ítems(23 ítems) con funcionamiento diferencial..

Concordancia entre procedimientos.

Para evaluar el nivel de acuerdo entre los procedimientos en cada una de las dos pruebas analizadas, se calculan las correlaciones entre detecciones. La tabla 4 muestra el resultado de este análisis. En ella pueden observarse el número de ítems con funcionamiento diferencial detectado por cada uno de los procedimientos, las correlaciones entre éstos y el número de ítems detectados simultáneamente.

Aptitud numérica: En esta prueba existe un nivel de concordancia absoluta entre los índices de área. Ambos detectan exactamente los mismos 7 ítems. Puede además observarse que la correlación que presentan con el índice de Lord es muy alta 0,831. De los 9 ítems que detecta este procedimiento 7 coinciden con AECS y AESS.

Este nivel de acuerdo entre los procedimientos derivados de la TRI no lo podemos generalizar al estadístico MH. Las correlaciones de $L\chi^2$, AECS y AESS con MH no son significativas, con valores de 0,359 en el primer caso y de 0,275 en el resto.

Comprensión verbal: En esta prueba, si bien no existen correlaciones perfectas entre ninguno de los procedimientos, hemos de decir que las técnicas derivadas de la teoría de respuesta a los ítems presentan correlaciones significativas al nivel 0,05, lo que indica un nivel de solapamiento alto entre procedimientos. Sin embargo las correlaciones que presentan con el estadístico MH no son significativas en ninguno de los casos (0,225(MH-Lord), 0,193 (MH-AES) y 0,279 (MH-AESS)).

Tabla 4. Concordancia ente procedimientos de detección del FDI.

	APTITUD NUMERICA				COMPRENSION VERBAL			
	Lord (9)	AECS (7)	AESS (7)	MH (6)	LORD (23)	AECS (17)	AESS (20)	MH (23)
LORD	1,000				1,000			
AES	0,831 (7)	1,000			0,515 (17)	1,000		
AESS	0,831 (7)	1,000 (7)	1,000		0,780 (20)	0,433 (15)	1,000	
MH	0,359 (4)	0,275 (3)	0,275 (3)	1,000	0,255 (19)	0,193 (15)	0,279 (17)	1,000

Causa del funcionamiento diferencial

Una vez aplicados los procedimientos estadísticos pertinentes para la detección del FD es imprescindible buscar las causas del mismo. En esta segunda parte del estudio entran en juego los procedimientos de juicio para evaluar la posible existencia de sesgo (Marascuilo y Slaughter, 1981).

Los autores que desde la TRI han analizado el problema de la equivalencia métrica definen las fuentes más comunes de divergencia en dos grandes grupos; las relacionadas con la traducción y las originadas por las diferencias en la relevancia de los ítems en los dos idiomas (Ellis, 1989; Ellis, Becker y Kimmel, 1993; Ellis y Kimmel, 1992; Hambleton, 1996;Hulin, 1987; Hulin y Mayer, 1986)

Con el fin de desechar la hipótesis de que las diferencias observadas se deben a un defecto en la traducción, que por otro lado acarrearía problemas de solución relativamente sencilla, se construye una prueba de jueces en la que se ha de valorar en una escala de 1 a 5 la idoneidad de la adaptación de cada uno de los ítems. Son cinco los jueces que intervienen en este estudio, 4 profesores bilingües de la Facultad de Psicología de San Sebastián y un traductor profesional ajeno a todo el proceso de adaptación original. Se les pide que juzguen los 30 ítems de la prueba de comprensión verbal y 11 de la prueba de aptitud numérica. Para cada uno de los ítems se presenta a los jueces la versión original y adaptada, solicitándoles que evalúen por separado la formulación y cada una de las alternativas de respuesta. La escala presentada es la siguiente: 1.- desacuerdo total 2.- problemas en la traducción 3.- cierta disparidad 4.- traducción aceptable 5.- concordancia total.

El índice de consistencia entre jueces evaluado con el alpha de Cronbach es de 0,97. El valor escalar menor no es inferior a 18. De las 246 presentaciones a valorar por cada uno de los jueces 230 obtienen valores escalares de 25. De este proceso concluimos la buena adaptación lingüística de las pruebas y descartamos la deficiencia en la traducción como posible causa de FD.

Aptitud numérica. En la prueba de aptitud numérica el componente verbal de los ítems se reduce al mínimo. Si revisamos el contenido de los 3 ítems (11, 17 y 19) que son catalogados como deficientes por todos los procedimientos de detección de FDI utilizados, vemos que la formulación verbal del enunciado se limita a instrucciones como **divide**o **multiplica** Dado que el nivel educativo de los sujetos a los que se administra la prueba es el mismo y que en el total de la prueba son 8 los ítems que exigen al sujeto labores de multiplicación o división y no todos ellos presentan FDI, parece del todo razonable rechazar la hipótesis de que la fuente de error esté en el distinto peso cultural de los ítems en ambas lenguas, y nos induce a buscarla

en los mismos procedimientos de detección, es decir en las tasas de error en que todo procedimiento estadístico puede incurrir.

Comprensión verbal. Esta prueba es eminentemente verbal y si bien las traducciones son literalmente correctas el porcentaje de ítems con funcionamiento diferencial detectado en contra de la muestra focal es desmesurado con cualquiera de los procedimientos utilizados. Intentando profundizar en la búsqueda de las fuentes de error y desechando la hipótesis de incorrección de la adaptación aceptamos la existencia de diferencia en la relevancia de los ítems en los dos idiomas.

En la búsqueda de razones que expliquen la existencia de sesgo, el punto de vista imperante en la actualidad es el de la multidimensionalidad. Desde esta perspectiva se considera que el ítem o la prueba que tiene sesgo, mide más de una dimensión en los dos grupos que se estudian y además las distribuciones multidimensionales de estos grupos son diferentes. Puede suceder que el ítem o la prueba que está sesgada mida en una población algo más que el factor único que mide en la otra. Si la distribución de los dos grupos en este segundo rasgo es distinta y los ítems son incapaces de percibirla, aparecerán como sesgados (Ackerman, 1992; Mellengergh, 1989).

Con este enfoque analizaremos de nuevo las estructuras factoriales de la prueba de comprensión verbal en las dos muestras. Si bien hemos indicado que ninguna de ambas alcanza el criterio de unidimensionaliad de Reckase, hemos de hacer hincapié en el hecho de que el índice propuesto por Lord es mayor en la muestra A que en la muestra D (11,16 frente a 4,5), lo que se traduce como mayor unidimensionaliad de la prueba en el primer caso que en el segundo.

Por otro lado si comparamos el grado de similitud factorial entre ambas muestras con el coeficiente de congruencia de Burt y Tucker (Burt, 1948) obtenemos un valor de 0,20 que nos obliga a rechazar la existencia de equivalencia. Aunque este recurso para la comparación de estructuras factoriales es bastante simple (siguiendo a Harman (1980) podemos decir que no se trata más que de un índice análogo a un coeficiente de correlación), en nuestro caso aporta evidencia confirmatoria de falta de unidimensionalidad de la prueba en la muestra Modelo D y apoya la hipótesis de existencia de sesgo y por tanto de falta de validez de constructo.

CONCLUSIONES

Las conclusiones de este trabajo podemos enmarcarlas en dos campos diferenciados aunque no por ello inconexos; uno asociado con la adaptación de pruebas psicopedagógicas y el otro relacionado con los procedimientos de detección del funcionamiento diferencial de los ítems.

Comenzando con este último punto, tendríamos que señalar en primer lugar las condiciones adversas en que se han utilizado los procedimientos de detección. La existencia de diferencias en la distribución de los grupos de referencia y focal junto al alto porcentaje de ítems con funcionamiento diferencial reduce la eficacia de los procedimientos y aumenta la tasa de errores Tipo I(Mazor, Clauser y Hambleton, 1994; Rogers y Swaminathan, 1993). Hemos de añadir además que el ajuste entre el modelo logístico de dos parámetros y los datos no es perfecto en la muestra D en ninguna de las pruebas analizadas, lo que obliga a evaluar con precaución los resultados de la aplicación de los procedimientos derivados de la TRI. El peso de estas desfavorables circunstancias puede haberse reflejado en el estudio de concordancia realizado.

En el caso de la prueba de aptitud numérica donde las diferencias de distribución en las dos muestras y el porcentaje de funcionamiento diferencial es menor que en la prueba de comprensión verbal, el estadístico MH que es el que menos número de detecciones presenta tiene correlaciones de 0,275 y 0,350 con los procedimientos derivados de la TRI. En el caso de la prueba de comprensión verbal donde los porcentajes de ítems con FD oscilan entre el 56% y el 76% en función del procedimiento utilizado, los resultados se repiten. Correlaciones significativas entre los procedimientos derivados de la TRI y correlaciones que no superan el valor de 0,279 con MH.

En trabajos anteriores en los que se han comparado estos estadísticos se llega a conclusiones de concordancia entre ellos (Budgell, Raju y Quartetti, 1995; Elosua, López, Artamendi y Yenes, manuscrito enviado para su publicación; Raju, Drasgow y Slinde, 1993; Hambleton y Rogers, 1989) que no podemos suscribir en este caso. Esta discrepancia puede deberse a las condiciones tan adversas de este estudio que no se dan en los trabajos citados en la misma cuantía (alto porcentaje de ítems con FDI, diferencias altamente significativas en las distribuciones de las poblaciones de referencia y focal y falta de ajuste del modelo logístico de dos parámetros).

La falta de un control externo sobre las condiciones de aplicación propias de todo trabajo empírico y las divergencias encontradas en la catalogación de ítems con funcionamiento diferencial por los procedimientos empleados, nos conducen a insistir en:

.- la necesidad de estudios de concordancia entre técnicas de detección como criterio externo de validación

.- la conveniencia de estudiar la estabilidad de los procedimientos de detección de FDI. La evaluación de estabilidad se podría llevar a cabo mediante la creación de submuestras aleatorias de la población focal. De este modo tendríamos dos grupos Referencia-Focal1 y Referencia-Focal2 que salvo errores muestrales habrían de ofrecer las mismas detecciones.

Los resultados de ambas técnicas de control servirían para apoyar estudios post hoc sobre las causas del funcionamiento diferencial, que separarían el FDI producido por la aplicación de las técnicas de detección de (errores tipo I), del FDI originado por la existencia de diferencias en el modo de responder al ítem entre sujetos con el mismo nivel de habilidad.

En cuanto al problema de la adaptación de pruebas psicopedagógicas, los resultados obtenidos contribuyen a reforzar la idea de necesidad de evaluación de la equivalencia métrica ya esbozada en trabajos anteriores (Elosua, López y Torres, 1997, 1999). Los estudios de equivalencia basados en la comparación de coeficientes de fiabilidad o de estructuras factoriales no aseguran en absoluto la igualdad psicométrica; esta ha de ser garantizada para cada uno de los ítems que componen la prueba, evaluando su posible funcionamiento diferencial como fase previa al estudio del sesgo.

Buscando una explicación al alto grado de funcionamiento diferencial y sesgo, y una vez desechados los problemas derivados de la traducción hemos de indagar en la distinta relevancia o familiaridad de los ítems en las dos lenguas.

La razón de la divergencia la encontramos en la falta correspondencia entre las dimensiones semánticas entre el castellano y el euskera, que ni la más perfecta de las traducciones asegura. Los términos o conceptos de una lengua poseen dimensiones semánticas que no necesariamente coinciden con su traducción literal a otra lengua, lo que crea una falta de concordancia semántica entre idiomas que es patente en las pruebas con fuerte componente verbal y se acentúa aún si cabe en una población infantil en período de adquisición de competencia lingüística plena. Sería preciso un estudio normativo que evaluara las dimensiones semánticas de las palabras para que la adaptación de pruebas avanzara más allá de la equivalencia literal, hacia una concordancia semántica que garantizase la igualdad en la familiaridad y significatividad de los términos. No es suficiente para el logro de la equivalencia métrica un buen conocimiento de las dos culturas. Mientras no se establezcan pautas normativas objetivas para el euskera basadas en la frecuencia de uso de las palabras como existen para el valenciano o el catalán (Algarabel, S; Ruiz, J,C.; Sanmartín, J, 1988; Nacher, Gotor y Algarabel 1998) las adaptaciones de pruebas psicológicas serán correctas desde un punto de vista meramente lingüístico pero no alcanzarán la equivalencia semántica y en consecuencia la equivalencia métrica.

ABSTRACT

Differential item functioning and bias in the adaptation of two verbal tests. This paper underlines the phases through which every psychological test adaptation should go. It is llustrated with a study of the adaptation of the Basque language (Euskera) to two tests containing

verbal components. In the first phase, which is of exploratory nature, the coefficients of reliability and factorial structures are compared and also the differential item functioning is detected. In the second phase, which is of confirmatory nature, the causes of DIF are analised in order to determine the existence or inexistence of bias. The DIF is computed using the Mantel-Haenszel statistic and three IRT based methods (Lord´s χ^2 and the signed and unsigned areas).The high percentage of DIF detected leads one to conclude that the adaptation, far from being straightforward linguistic translations, must take into account semantic dimensions.

Key words: differential item functioning, bias, metric equivalence, adaptation, item response theory, signed area, unsigned area, Mantel-Haenszel, Lord's χ^2 .

REFERENCIAS

Ackerman,T.A.(1992) Didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of educational measurement*, 29(1), 67-91.

Algarabel, S.,Ruiz.J.C. y Sanmartín,J.(1988). The university of Valencia's computerized word pool. *Behavior research methods, instruments and computers*, 20, 398-403.

Baker, F.B. (1994) EQUATE2: Computer program for equating two metrics in item response theory [Computer program]. Madison: University of Wisconsin, Laboratory of Experimental Design

Bontempo, R.(1993). Translation fidelity of psychological scales: An item response theory analysis of an individualism-collectivism scale. *Journal of cross-cultural psychology*, *24*(2), 149-167.

Brislin, R.W.(1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, 1(3), 185-216.

Budgell, G.R., Raju, N.S. y Quarteti, D.A.(1995) Analysis of differential item functioning in translated assessment instruments. *Applied psychological Measurement*, 19(4), 309-321.

Burt,C.L. (1948) The factorial study oftemperamental traits. *British journal of psychplogy*, 1, 178-203.

Candell, G.L. y Drasgow, F.(1988): An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied psychological measurement, 12*(3), 253-260.

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*, 1(4), 417-440.

Casagrande,J.B. (1954): The ends of translation. *International Journal of American Linguistics*, XX, 335-340.

Drasgow, F. (1984). Scrutinizing psychological test: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, 95(1),134-35.

Drasgow,F. y Hulin, C.L. (1987). Cross-cultural meassurement. *Revista interamericana de psicología/Interamerican Journal of Psychology*, 21(1,2), 1-24.

Drasgow, F. y Lissak, R.I. (1983) Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied psychology*, 68(3), 363-373.

Ellis, B.B.(1989) Differential item functioning: implications of tests translation. *Journal of applied psychology*, 74(6), 912-921.

Ellis,B.B. (1991) Item response theory: a tool for assessing. *Bulletin of the international test commission*, 18, 33-51.

Ellis, BB., Becker.P. y Kimmel, H.D.(1993). An item response theory evaluation on an english version of the Trier Personality Inventory (TPI). *Journal of Cross-cultural psychology*, 24(2), 133-148.

Ellis, B.B., Minsel, B. y Becker, P. (1989) Evaluation of attitude survey translations: an investigation using item response theory. *International journal of psychology*, 24, 665-684.

Elosua, P., López, A., Artamendi, J.A. y Yenes, F. Funcionamiento diferencial de los ítems en la aplicación de pruebas psicológicas en entornos bilingües. En revisión.

Elosua, P. , López, A. y Torres, E. (1997, Septiembre) *Adaptación al euskera de una prueba verbal. Estudio del funcionamiento diferencial de los ítems. Concordancia entre los procedimientos Mantel-Haenszel, Logit iterativo, Regresión logística iterativa y SIBTEST.* Comunicación presentada en el Congreso de Metodología de las Ciencias Sociales y del Comportamiento. Sevilla.

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

Feldt, L.S.(1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two test. *Psychometrika, 34*, 363-373.

Ferrando, P.J.(1996). Evaluación de la unidimensionalidad de los ítems mediante análisis factorial. *Psicothema, 8*(2), 397-410

Fidalgo, A.M.(1994). MHDIF: A computer program for detecting uniform and nouniform differentzial item functioning with the Mantel-Haenszel procedure.[Computer program] Dpto. Psicología, Universidad de Oviedo.

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.207-238). Madrid: Universitas, S.A.

Hambleton, R.K.(1993). Translating achievement test for use in cross-national studies. *European Journal of Psychological Assessment, 9*(1), 57-68.

Hambleton, R.K. y Bollwark,J. (1991). Adapting test for use in different cultures:technical issues and methods. *International Test Bulletin*, 32/33, 3-32.

Hambleton, R.K. y Swaminathan, H.(1985) *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff

Harman, H.H. (1980). *Análisis factorial moderno.* Madrid:Saltés

Holland, P.W. y Thayer, D.T.(1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer y H.J. Braun (eds.), Test validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Hulin, C.L. (1987). A psychometric theory of evaluations of item scale translations. *Journal of cross-cultural Psychology*, 18(2),115-142

Hulin, C.L., Drasgow, F. y Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology*, 67(6), 818-825).

Hulin, C.L., Drasgow, F. y Parsons, C.K. (1983) *Item response theory: Application to psychological measurement*. Homewood, Illinois: Dow Jones/Irwin.

Hulin, C.L. y Mayer,L. (1986) Psychometric equivalence of a translation of the job descriptive index into hebrew. *Journal of applied psychology*, 71(1), 83-94.

Kim, S.H. y Cohen, A.S. (1991) A comparison of two area measures for detecting differential item functioning. *Applied psychological measurement*, 15(3), 269-278.

Kim S.H. y Cohen, A.S. (1992a) Effects of linking methods on detection of DIF. *Journal of educational measurement, 29*(1), 51-66

Kim S.H. y Cohen, A.S. (1992b). IRTDIF: A computer program for IRT differential item functioning analysis [Computer Program] University of Wisconsin-Madison.

Kolen, M.J.(1981). Comparation of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18*, 1-11.

Linn, R.L. y Harnisch, D.L. (1981) Interactions between item content and group membership on achievement test items. *Journal of Educational measurement*, 18(2), 109-118.

López Pina, J.A.(1995). *Teoría de la respuesta al ítem: fundamentos*. Barcelona:PPU

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

Mantel, N. y Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, 22, 719-748.

Marascuilo, L.A. y Slaughter, R.E.(1981). Statistical procedures for identifying posible sources of item bias based on χ^2 statistics. *Journal of Educational Measurement*, 18(4), 229-248.

Mazor, K.M., Clauser, P.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*, 284-291.

Mellenbergh,G.J.(1989) Item bias and item response theory. *International journal of educational research,* 13, 127-143.

Mislevy, R.J. y Bock, R.D.(1990). BILOG·3: Item analysis and test scoring with binary logistic models.[Computer program]. Mooresville, IN: Scientific software.

Muñiz, J.(1990). *Teoría de respuesta a los ítems*. Madrid. Pirámide.

Nácher, M.J., Gotor, A. y Algarabel, S. (1998) Traducciones equivalentes en catalán y castellano de 1533 palabras y sus valores normativos en concreción, familiaridad y significatividad. *Psicológica 19*, 1-26.

Park, D.G. y Lautenschlager, G.J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied psychological measurement, 14*(2),163-173.

Pine, S.M.(1977). Applications of item response theory to problem of test bias. En D.J. Weiss(Ed.), Applications of computerized adaptive testing (pp.37-43); (Research Report Nº77-1). Minneapolis: University of Minnesota.

Raju, N.S.(1988) The area between two item characteristic curves, *Psychometrica*, 53(4), 495-502

Raju, N.S.(1990) Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14(2), 197-207.

Raju, N.S. Drasgow, F. y Slinde, J.A.(1993): An empirical comparison of the area methods, Lord's chi square test, and the Mantel-Haenszel technique for

assessing differential item functioning. Educational and psychological measurement, 53, 301-304

Reckase, M.D.(1979): Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics*, 4, 207-230.

Rogers, H.J.y Swaminathan, H. (1993). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied psychological measurement,* 17(2), 105-117.

Rudner, L.M., Getson, P.R. y Knight, D.L. (1980). A montecarlo comparison of seven biased item detection techniques. *Journal educational measurement,* 17(1), 1-10.

Scheuneman, J.D.(1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16*(3), 143-152.

Shealy, R. y Stout, W.(1993). An item response theory model of test bias and differential test functioning. En W.P. Holland y H. Wainer (Eds.), *Differential item functioning* (pp. 197-240). Hillsdale, NJ: Lawrence Erlbaum.

Stocking, M.L. y Lord, F.M. (1983) Developing a common metric in item response theory. *Applied psychological measurement, 7*(2), 201.210.

Thissen, D. y Wainer, H.(1982). Some standards errors in item response theory. *Psychometrika, 47*(4), 397-412.

Yuste, C. (1988). *BADYG-E*. Madrid. Ciencias de la educación preescolar y especial.

(Revisión aceptada: 22/3/99)