Psicothema
2000. Vol. 12, nº 2, pp. 201-206

ISSN 0214 - 9915 CODEN PSOTEG

Idioma de aplicación y rendimiento en una prueba de comprensión verbal

Paula Elosua Oliden, Alicia López Jáuregui y Josu Egaña Makazaga
Universidad del País Vasco

En este trabajo se intentan evaluar los efectos que tiene sobre el rendimiento y sobre la validez tanto la administración de un test en un idioma que no coincide con el idioma dominante de los sujetos destinatarios como su adaptación al euskera. Para ellos se comparan los resultados obtenidos en una prueba de comprensión verbal por una muestra de referencia con dos muestras sometidas a las condiciones mencionadas. El alto porcentaje de funcionamiento diferencial evidencia la falta de validez en cualquiera de las dos condiciones evaluadas.

Administration language and differential item functioning. The purpose of the present research was to evaluate the effects in score and validity, of one test administration in a non dominant language for the subjects, also the score and validity of the instrument translated to the basque language. The results in a verbal comprehension test were compared in relation to the reference and two focal groups. The high percentage of differential item functioning show a lack of validity of the test in anyone of the two evaluated conditions.

Uno de los problemas más representativos de la Teoría Clásica de Tests es la falta de invarianza de las mediciones con respecto a la población y a las propiedades de los instrumentos de medida (Muñiz y Hambleton, 1992). Esto supone que la interpretación de las puntuaciones está siempre condicionada. Por ello, dentro del estudio de validez a que se somete todo test de evaluación y diagnóstico psicológico es necesario precisar de modo conciso junto con las características de la población destinataria, las condiciones de aplicación que lo optimicen. Esta circunstancia obliga a que siempre que se produzca alguna alteración sobre las especificaciones originales, bien en la población, bien en la administración, sea necesario volver a estudiar las propiedades psicométricas con el fin de confirmar la validez. La APA, AERA y NCME reconocen explícitamente este hecho al afirmar en los estándares publicados en 1985:

> «Cuando se hacen cambios en el formato del test, modo de aplicación, instrucciones, idioma o contenido, el usuario debería de revalidar el test para las nuevas condiciones, o tener argumentos que apoyen que no es necesario o posible una validación adicional» (estándar 6.2. pág.41)

Las comunidades donde existe contacto de lenguas son especialmente proclives a este tipo de cambios de formato. Puede alterarse el idioma original de la prueba adaptándolo a un segundo idioma, o puede utilizarse en la administración del instrumento un idioma que no coincide con el idioma dominante de los sujetos destinatarios. Tanto en una situación como en la otra se están vulnerando las condiciones originales en las que se han fundamentado las características métricas, lo que exige, de acuerdo con el estándar citado, la revalidación de las nuevas situaciones.

En la Comunidad Autónoma del País Vasco no existen prácticamente pruebas ni adaptadas ni creadas en euskera, por lo que se convierten en práctica común tanto la traducción lingüística como la administración de tests a sujetos en su segunda lengua. Estos hábitos conllevan la aceptación de supuestos, no probados empíricamente, que implican la existencia de equivalencia métrica entre las puntuaciones obtenidas de uno u otro modo.

En este contexto general nuestra investigación evalúa estas asunciones que formalizamos a través de las siguientes hipótesis:

1. La traducción literal de las pruebas de medición psicopedagógica no garantiza en sí misma la equivalencia métrica de las puntuaciones.

2. La administración de un instrumento de medición psicológica a sujetos bilingües en un idioma que no coincide con el de su lengua materna y el de su instrucción, no garantiza la equivalencia métrica de las puntuaciones.

El presente estudio analiza las implicaciones teóricas y prácticas del uso de estas estrategias de evaluación en una población infantil que por estar inmersa en pleno proceso de adquisición y dominancia lingüística se ve especialmente afectada por cualquier influjo idiomático.

La verificación de estas dos hipótesis exige un análisis de validez, que examine por un lado, la equivalencia entre los coeficientes de fiabilidad y estructuras factoriales, y por otro, la existencia de una relación idéntica en las dos poblaciones entre cada uno de los ítems que componen la prueba y el rasgo medido (Drasgow, 1984). Es decir, es necesario garantizar la equivalencia métrica de modo que la probabilidad de responder correctamente a un ítem

Correspondencia: Paula Elosua Oliden
Facultad de Psicología
Universidad del País Vasco
20009 San Sebastián (Spain)
E-mail: pspelolp@sc.ehu.es

sea independiente del grupo de pertenencia, y esté sólo en función del nivel del sujeto en el rasgo; en caso contrario concluiríamos la existencia de funcionamiento diferencial del ítem (FDI).

El análisis se efectúa haciendo uso de la tecnología derivada de la Teoría de Respuesta al Item para la detección del funcionamiento diferencial de los ítems (FDI), pues creemos que la definición de equivalencia métrica aportada por Drasgow se ajusta perfectamente a los dos objetivos perseguidos. De la existencia de equivalencia entre los coeficientes que evalúan los tests como totalidad, no se puede concluir equivalencia psicométrica entre ítems, condición necesaria para garantizar la validez de las interpretaciones.

La utilización del concepto de funcionamiento diferencial del ítem ha resultado fructífera en el proceso de validación que ha de seguir toda adaptación de pruebas, y así los atestiguan los trabajos de 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 (1999a), Elosua y López (1999b), Hulin (1987), Hulin, Drasgow y Komocar (1982), Hulin, Drasgow y Parson (1983), Hulin y Mayer (1986).

En el caso de los estudios que evalúan las implicaciones de la administración de pruebas a sujetos bilingües, podemos diferenciar dos tipos de trabajos en función de la metodología utilizada. Los que recurren y se benefician del concepto de funcionamiento diferencial y aquellos que derivan sus conclusiones de la comparación de rendimientos tras la administración de pruebas. Dentro de estos últimos podríamos citar los trabajos de Artamendi (1994), Palmer (1972), Kelly Tenezakis y Huntsman (1973) y Collison (1975), de los que se desprende que si la lengua materna no es la lengua de instrucción, no supone ninguna ventaja la aplicación de pruebas en dicha lengua. Por otro lado en los trabajos que incluyen el concepto de FDI (Hulin, Drasgow y Komocar, 1982; Candell y Hulin, 1987) el objetivo es diferente. No pretenden evaluar la diferencia de rendimiento ni la validez de las inferencias, en su lugar la finalidad última es analizar la calidad de una adaptación lingüística en la que los bilingües, en estos casos sujetos adultos, son el instrumento que permite valorar la idoneidad de una traducción, en virtud de su conocimiento de los dos idiomas.

La metodología utilizada en este trabajo, parte de una descripción general del rendimiento en función de las condiciones citadas (adaptación lingüística y administración en la segunda lengua), que ampliaremos valorando las diferencias existentes por grupos de edad o curso. Continuaremos con una evaluación psicométrica que incluye un estudio de la estructura factorial, ajuste a los modelos logísticos y detección del funcionamiento diferencial de los ítems con el estadístico Mantel-Haenszel (Holland y Thayer, 1988) y el chi-cuadrado de Lord (1980).

Método

Sujetos

La muestra está formada por 1.999 niños que estudian en 4º, 5º y 6º de enseñanza primaria, repartidos en 11 centros escolares de la Comunidad Autónoma del País Vasco; a Guipúzcoa pertenecen 959 sujetos, a Bizkaia 493 y a Alava 545. Para la administración de los cuestionarios a los niños euskaldunes se contó con la colaboración de la Federación de Ikastolas que envió una carta a cada uno de los centros implicados solicitando su colaboración (las ikastolas son centros concertados en los que la docencia se imparte exclusivamente en euskera, excepto en la asignatura de Lengua Castellana). Los sujetos castellanoparlantes provienen de dos centros de enseñanza, uno público y el otro privado.

Para alcanzar los objetivos propuestos, la muestra total se divide en dos submuestras. La primera de ellas a la que denominamos genéricamente CC esta formada por sujetos monolingües castellanoparlantes y será la muestra de referencia en todos los análisis que vayamos a efectuar. Las características de este grupo se ajustan al grupo normativo en que se ha basado el estudio psicométrico original.

En el polo lingüísticamente opuesto situaríamos a la segunda muestra, formada por sujetos que además de poseer como lengua materna el euskera reciben la instrucción en este mismo idioma (Modelo D de enseñanza bilingüe). Dentro de este segundo grupo volvemos a diferenciar dos submuestras según el idioma de administración de la prueba. La prueba administrada en castellano da origen al grupo EC (grupo euskaldun ejecuta la prueba original en castellano). La prueba adaptada al euskera por su parte forma el grupo EE (grupo euskaldun ejecuta la prueba en euskera). La selección de estas dos submuestras ha sido aleatoria dentro de cada ikastola y curso. La comparación a que da lugar este último conjunto con la muestra CC se ajusta al diseño definido por Hambleton (1993, 1996) para el estudio de la calidad psicométrica de las adaptaciones de instrumentos de medida.

Tenemos por tanto un grupo de referencia y dos grupos focales con los que podemos cotejar las siguientes diferencias CC-EC, CC-EE y EC-EE.

El instrumento que vamos a analizar es la prueba de Comprensión Verbal (CV) perteneciente a la batería de aptitudes diferencial y general (Yuste, 1988) en su versión elemental. Es una prueba que consta de 30 ítems dicotómicos con 5 alternativas de respuesta. El coeficiente de fiabilidad referido en el manual y calculado con el método de dos mitades y la corrección de Spearman-Brown tiene un valor de 0,84 (no se ofrece el coeficiente de consistencia interna). Los ítems se clasifican en sinónimos, antónimos, analogías verbales y búsqueda de la definición más exacta de un concepto. La puntuación obtenida en esta prueba indicaría conocimiento de vocabulario y facilidad en el reconocimiento de relaciones verbales analógicas.

La adaptación de la prueba al euskera ha seguido las pautas que marca la retrotraducción (Brislin, 1970) y que se han adoptado en trabajos anteriores (Elosua, López y Torres, 1999a). 1. Un grupo de sujetos bilingües traduce la prueba al euskera; 2. Otro grupo independiente del anterior retrotraduce la prueba al castellano 3. Se analizan las divergencias con la ayuda de un profesor de enseñanza primaria, buscando ante todo la acomodación de contenidos al nivel de los sujetos destinatarios.

Resultados

La tabla 1 muestra los primeros estadísticos descriptivos, donde puede apreciarse que el rendimiento mayor es el obtenido por la muestra de referencia o muestra CC, y el menor se asocia con los sujetos a los que se administra la prueba en euskera. El ANOVA de un factor en el que se evalúa el efecto de la variable «idioma de aplicación» produce un valor significativo ($F_{2,1996}$=534,89; p<0,001) que refleja la influencia determinante de este factor sobre la puntuación final. La ordenación de los grupos en función del rendimiento daría la siguiente clasificación: 1. CC; 2. EC y 3. EE, con una diferencia de medias entre las dos muestras extremas de 8,37 puntos.

Tabla 1
Descripción de las muestras y consistencia interna

	N	$\bar{X}$	Sx	α
CC	545	20,67	4,77	0,78
EE	933	12,30	4,57	0,73
EC	521	15,00	5,33	0,81

Diferencias en el rendimiento por idioma y curso

Un análisis más detallado nos lleva a evaluar las diferencias existentes en el rendimiento en función del curso y del idioma de administración.

La tabla 2 muestra un estudio post hoc de todas las diferencias de rendimientos entre los grupos en los tres cursos y su significación (α= 0,05) evaluada con la prueba de Tukey-b. Con relación a la comparación entre la muestra de referencia CC y la focal EC podemos apreciar que la diferencia máxima está asociada al curso 4º (8,20; p<0,0001) y que se mantiene prácticamente constante en los cursos 5º y 6º (5,35; p<0,0001 y 5,56;p<0,0001). Si la comparación la efectuamos entre los dos grupos focales (EC y EE) apreciamos que, siendo el rendimiento de estas muestras similar y estadísticamente no significativo en el curso 4º (la diferencia de medias es de 0,45; p<0,9), estas diferencias aumentan progresivamente en los cursos 5º y 6º alcanzando los valores de 2,36 (p<0,0001) y 3,65 (p<0,0001) respectivamente a favor de la muestra euskaldun a la que se administra la prueba en castellano.

En la comparación del rendimiento entre las muestras más polarizadas CC y EE sale beneficiada la muestra de referencia. Es en este par donde encontramos sistemáticamente las diferencias más acusadas, 8,66 (p<0,0001), 7,72 (p<0,0001) y 9,21 (p<0,0001) para los cursos 4, 5 y 6 respectivamente.

Tabla 2
Comparaciones múltiples por idioma y curso (* p<0,05)

		4	5	6
CC	EC	8,20*	5,35*	5,56*
CC	EE	8,66*	7,72*	9,21*
EC	EE	0,45	2,36*	3,65*

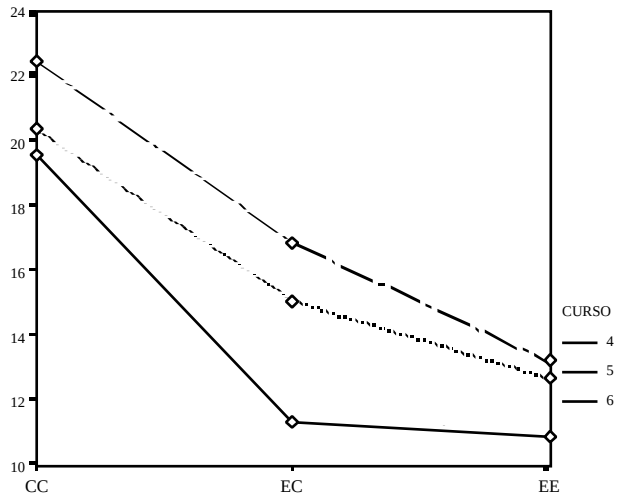

Gráfico 1. *Rendimiento por idioma de aplicación y curso*

La comparación de los coeficientes de consistencia interna entre muestras se formaliza con el estadístico de Feldt (1969). Los valores alcanzados en las comparación CC-EC (1,15), CC-EE (0,814) y EC-EE (1,42) nos conducen a aceptar la hipótesis nula de igualdad de α.

Unidimensionalidad

La aplicación más productiva de un modelo de teoría de respuesta al ítem unidimensional se obtiene con la presencia de esta condición en los datos. Para evaluar este requisito se someten a un análisis de componentes principales las matrices de correlaciones tetracóricas obtenidas en cada una de las muestras. Los resultados de este análisis se ofrecen en la tabla 3.

Tabla 3
Autovalores y % de varianza explicada

	CC		EC		EE	
factor	Autovalor	%Varianza	Autovalor	%Varianza	Autovalor	%Varianza
1	7,04	24,29	7,51	25,03	5,47	18,24
2	1,84	6,36	1,99	6,63	2,11	7,04
3	1,49	5,13	1,57	5,25	1,53	5,12

Puede apreciarse que en las muestras de referencia y en la focal EC se supera el criterio de unidimensionalidad de Reckase (1979) según el cual el primer factor ha de explicar el 20% de la varianza. En la muestra EE el primer factor extraído explica el 18,24% de la variabilidad total. Por otro lado, los gráficos de sedimentación apoyan la presencia de un factor dominante (Tatsuoka, 1988) en cada una de las muestras, si bien es cierto que la diferencia entre el primer factor y el segundo es menor en la muestra EE que en el resto. En este punto señalaremos que si la matriz a factorizar es la matriz de correlaciones phi el porcentaje de varianza explicada por el primer factor en cada una de las muestras sería de 14,4% (CC), 16% (EC) y 12,3% (EE). Tal divergencia de resultados puede deberse a la existencia de un rango de valores en los índices de dificultad de los ítems que atenuarían los coeficientes phi produciendo evidencia espuria de multidimensionalidad (Ferrando, 1994; Collins, McCormick y Zatkin, 1986)

Estimación de los parámetros

Los parámetros de los ítems se estiman independientemente en cada una de las muestras utilizando para ello el procedimiento de estimación de máxima verosimilitud marginal implementado en el programa BILOG3 (Mislevy y Bock, 1990). Se ajustan los modelos logísticos de dos y tres parámetros. La muestra CC no presenta valores significativos para ninguno de los ítems (p<0,01). En la muestra EC dos ítems (1-18) obtienen valores χ^2 con una significación p<0,01 en el modelo logístico de dos parámetros. Si evaluamos el modelo de tres parámetros el ítem 1 tiene una χ^2 significativa (χ^2=14,4; p<0,0063). En la muestra EE los ítems 13, 16 y 22 alcanzan esta significación tanto en el modelo de dos parámetros (χ^2=25,1; p<0,029; χ^2=25,7; p<0,0024; χ^2=23,5; p<0,0052) como en el de tres parámetros (χ^2=28,2; p<0,0009; χ^2=27,2; p<0,0014; χ^2=26,3; p<0,0019). Los valores chi-cuadrado para la evaluación del ajuste total no son significativos en las muestras CC

y EC en ninguno de los modelos aplicados, mientras que en la muestra EE estos valores son p≤0,0001 en el modelo de dos parámetros y de p≤0,000001 en el modelo de tres parámetros. El principio de parsimonia unido a estas circunstancias hacen que optemos por la elección del modelo logístico de dos parámetros.

Funcionamiento diferencial de los ítems

Antes de la detección del funcionamiento diferencial de los ítems se equiparan las métricas de los grupos a comparar con el procedimiento de la curva característica sugerido por Stocking y Lord (1983) e implementado en el programa EQUATE2 (Baker, 1994). El test de anclaje para cada una de las comparaciones esta formado por los ítems que tienen valores χ^2 no significativos (p<0,01). Para evaluar el FDI aplicamos dos procedimientos; uno derivado directamente de la teoría de respuesta a los ítems (χ^2 de Lord (1980)) y el otro basado en el estudio de tablas de contingencia, el estadístico Mantel-Haenszel (Holland y Thayer, 1988).

Lord (1980) 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. En el caso de que los parámetros de los ítems coincidan, salvo errores de muestreo, las curvas características derivadas de ellos serán idénticas y se concluirá ausencia de funcionamiento diferencial. En la aplicación iterativa de este procedimiento (Candell y Drasgow, 1988) los resultados se estabilizan en dos fases. Los cálculos se efectúan con el programa IRTDIF (Kim y Cohen,1992)

Mantel-Haenszel. Es un procedimiento para el estudio de las tablas de contingencia que compara las respuestas dadas a un ítem por sujetos que perteneciendo a distintas poblaciones están agrupados en un mismo nivel de puntuación. Se contrasta la hipótesis nula 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. Los cálculos se efectúan con el programa MHDIF (Fidalgo, 1994), que implementa un procedimiento en dos fases para la purificación del criterio.

Aplicados ambos procedimientos a las comparaciones CC-EC y CC-EE se obtienen los resultados visibles en la tabla 4, donde el asterisco esta asociado con un nivel de significación de 0,05. No parece preceptivo llevar a cabo la comparación EC-EE dado que en ambos casos se vulneran las condiciones originales de administración y por lo tanto queda cuestionada su validez, objetivo de este trabajo.

Existe un acuerdo total en el número de detecciones mostradas por ambos procedimientos en las dos comparaciones. Son 15 los FDI referidos a la comparación CC-EC y 23 los asociados a la evaluación de CC-EE. Puede apreciarse que el porcentaje es en el primer caso del 50% y en el segundo alcanza el 76%.

Esta igualdad cuantitativa sin embargo no es cualitativa; No se identifican los mismos ítems. No existe un solapamiento del 100% en la catalogación de los ítems que presentan funcionamiento diferencial. En la situación más favorable, la que corresponde a la comparación, CC-EC, son 13 el número de solapamientos, lo que se refleja en una correlación phi entre las variables dicotómicas detección-no detección de 0,773 que resulta estadísticamente significativa (p<0,01). En la comparación CC-EE los 14 solapamientos no dan lugar a un coeficiente de correlación significativo (r= 0,255; p<0,174)

La causa de estos resultados en apariencia totalmente divergentes las podemos buscar en las diferencias entre las distribuciones empíricas de las muestras y en el porcentaje de funcionamiento diferencial. Existe una relación inversa entre la eficacia de los procedimientos de detección de funcionamiento diferencial de los ítems y las condiciones mencionadas, de modo que a mayor diferencia entre las distribuciones de habilidad de los grupos de referencia y focal y a medida que aumenta el porcentaje de ítems con funcionamiento diferencial se produce un decremento en la efectividad de estas técnicas (Rogers y Swaminathan, 1993; Mazor, Clauser y Hambleton, 1994). En nuestro caso en la comparación CC-EC existe un nivel alto de concordancia entre el χ^2 de Lord y el estadístico Mantel-Haenszel que ha sido también encontrada en otros trabajos empíricos (Hambleton y Rogers, 1989; Raju, Drasgow y Slinde, 1993; Budgell, Raju y Quarteti, 1995; Elosua, López, Artamendi y Yenes, manuscrito enviado para su publicación). No podemos sin embargo mantener la misma afirmación para la comparación CC-EE, donde la diferencia entre las distribuciones es de prácticamente dos desviaciones estándar y el porcentaje de funcionamiento diferencial es del 76%.

Una vez detectados los ítems con funcionamiento diferencial es necesario ir más allá y buscar las razones del mismo (Maras-

Tabla 4
Funcionamiento diferencial de los ítems (* p<0,05)

Ítem	CC-EC		CC-EE	
	Lord χ^2	MH D-DIF	Lord χ^2	MH D-DIF
1	2.31	0.75	1.93	-1.48*
2	1.46	0.22	26.67*	-3.62*
3	4.76	-1.20*	6.35*	-0.85*
4	20.57*	-1.78*	22.38*	1.33*
5	4.97	0.33	4.79	0.43
6	40.03*	1.94*	20.84*	0.36
7	6.01*	-0.21	22.73*	2.15*
8	2.41	-0.37	69.01*	-4.21*
9	2.61	-0.70	22.35*	-1.75*
10	26.70*	1.64*	54.28*	2.24*
11	0.72	-0.47	12.60*	-0.61
12	15.59*	-1.93*	10.14*	-2.13*
13	11.85*	-1.89*	7.95*	0.66
14	75.22*	2.81*	51.87*	1.29*
15	25.88*	1.49*	12.91*	-4.29*
16	1.54	-0.13	0.61	0.19
17	11.30*	1.79*	11.19*	-1.21*
18	2.19	-0.26	53.98*	1.60*
19	7.86*	0.64	106.67*	2.22*
20	15.54*	-2.03*	10.27*	-6.60*
21	0.40	-0.06	15.64*	0.85*
22	1.04	0.08	63.75*	1.84*
23	9.98*	0.92*	76.05*	1.70*
24	4.86	-1.26*	0.60	-0.47
25	14.87*	-2.24*	1.65	-1.13*
26	19.17*	1.37*	9.24*	0.42
27	5.41	0.10	4.51	-2.00*
28	7.87*	-1.72*	1.40	-1.73*
29	2.44	0.24	185.39*	3.68*
30	0.97	0.21	136.81*	2.95*
total	15	15	23	23
	A=1.6 K=-1.25		A=0.80 K=-1.93	
	A=1.03 K=-1.28		A=0.83 K=-2.07	

cuilo y Slaughter, 1981). Dentro de las posibles causas podríamos citar en primer lugar problemas derivados directamente del proceso de traducción (Ellis, Becker y Kimmel, 1993; Hambleton, 1996). Si bien esta podría ser la explicación del FDI asociado a la comparación CC-EE, en un estudio anterior (Elosua y López, 1999b) en el que se evalúa la calidad lingüística de la adaptación, se concluye su idoneidad y se desecha la hipótesis de existencia de errores lingüísticos en la traducción como fuente de funcionamiento diferencial.

Las diferencias en la relevancia cultural o significado del ítem y las diferencias de nivel culturales (Ellis, Becker y Kimmel, 1993) son también esgrimidas como fuente de funcionamiento diferencial. En un siguiente estadio del análisis cabría un estudio pormenorizado de cada uno de los ítems con funcionamiento diferencial que estudiara entre otros aspectos su sentido y tipo (uniforme, no uniforme, mixto). Sin embargo el objetivo de este trabajo se limita únicamente a probar la falta de validez de dos prácticas incorrectas y no obstante de uso generalizado en el entorno psicopedagógico de la Comunidad Autónoma del País Vasco.

Conclusiones

Los resultados obtenidos apoyan las hipótesis planteadas. El alto porcentaje de funcionamiento diferencial de los ítems evidencia una falta de validez que imposibilita cualquier interpretación de las puntuaciones basadas tanto en la traducción literal como en la aplicación de pruebas en castellano a sujetos euskaldunes que cursan estudios en el modelo D. Todos los análisis estadísticos realizados en este trabajo subrayan esta circunstancia.

Por otro lado si los que cotejamos son los dos grupos euskaldunes (EC y EE), el menos favorecido es el EE. El rendimiento de la muestra EC es significativamente mejor en los cursos 5º y 6º. Este resultado, que a primera vista parece contradictorio, podemos explicarlo con el porcentaje de funcionamiento diferencial de los ítems detectado en cada una de las comparaciones. El porcentaje es mayor cuando se cotejan los grupos CC-EE que cuando se confrontan las submuestras CC-EC (un 76% frente a un 50%). Esta circunstancia adversa tiene una relación directa sobre la puntuación final y por tanto sobre el rendimiento.

Si partimos de la definición de lo que es y pretende medir el test de Comprensión Verbal parece lógico llegar a esta conclusión. Dado que el objetivo de la prueba es dar una medida de conocimiento de vocabulario castellano en sujetos con este mismo idioma dominante, su aplicación en una muestra euskaldun (EC) medirá conocimiento de vocabulario en una segunda lengua, y en la muestra EE medirá conocimiento del euskera con ítems o términos que por estar traducidos literalmente no tienen por qué mantener el contenido semántico original. En ambos casos se produce una alteración de los objetivos generales de la prueba que propicia la transformación de las propiedades psicométricas de los ítems.

A la vista de estos resultados podría parecer más conveniente la administración a niños euskaldunes de pruebas en castellano frente la aplicación de pruebas adaptadas, sin embargo, hemos de recordar que en ambos casos la vulneración de las condiciones psicométricas ha creado una falta de validez de constructo operacionalizada en este caso por un elevado porcentaje de funcionamiento diferencial. Por lo tanto la aparente mejora del rendimiento no viene avalada por una equivalencia métrica, (Drasgow, 1984) lo que se traduce en la imposibilidad de asumir la igualdad entre el significado de las puntuaciones obtenidas en las muestras comparadas. Ante este hecho parece aconsejable la utilización del euskera en la evaluación de niños euskaldunes. Esto supone que el proceso de adaptación de instrumentos de medida ha de ser guiado en todo momento por el criterio de equivalencia psicométrica. No es suficiente que una adaptación sea lingüísticamente perfecta, sino que además ha de tener en cuenta los criterios internos propios de un idioma tales como el control de la tipografía, ortografía, morfología, léxico, corrección gramatical, adecuación y coherencia, así como las distintas dimensiones semánticas que garanticen la equivalencia en el grado de familiaridad y significatividad. Sólo la consideración de estas dimensiones en el proceso de adaptación, podrá garantizar pruebas sin funcionamiento diferencial, condición indispensable para que dos instrumentos de medida puedan ser considerados equivalentes.

Agradecimientos

Este trabajo ha sido subvencionado por la Universidad del País Vasco UPV 109.231-HA093/96.

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.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington, DC.American Psychological Association

Artamendi, J.A. (1994) Rendimiento de bilingües en test de aptitudes según lengua/s de presentación. En I.Idiazabal y A.Kaifer (Eds.),*Eficacia educativa y enseñanza bilingüe en el País Vasco* (pag.89-116) Vitoria: Instituto Vasco de Administración Pública

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.

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.

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.

Collins, McCormick y Zatkin (1986) Factor recovery in binary data sets a simulation. *Multivariate Behavioral Research 21*, 377-391

Collison, G.(1974) Concept formation in a second language: a study of gahaniaian school children. *Harvard Educational Review*, 44, 441-457.

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. y López.A(1999b) Funcionamiento diferencial de los ítems y sesgo en la adaptación de dos pruebas verbales. *Psicológica* 20(1),23-40

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

Elosua,P., López,A., Artamendi,J.A. y Yenes,F. (manuscrito enviado para su publicación) Funcionamietno diferencial de los ítems en la aplicación de pruebas psicolo gicas en entornos bilingües.

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.(1994) El problema del factor de dificultad: una revisión y algunas consideraciones prácticas. *Psicológica*, 15,2,275-283.

Fidalgo,A.M.(1994) MHDIF: a computer program for detecting uniform and nonuniform differential item functioning with the Mantel-Haenszel procedure [computer program] Dpto. Psicología, Universidad de Oviedo.2

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.(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. 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.

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. J*ournal 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.

Kelly, M., Tenezakis, M. Y Huntsman, R. (1973) Some unusual conservation behavior in children exposed to two culures. *British journal of educational psychology*, 43, 171-182

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

Lord, F.M.(1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

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

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. Y Hambleton, R.K. (1992) Medio siglo de teoría de respuesta a los ítems. *Anuarío de psicología.* 52, 41-66.

Palmer, M.(1972) Effects of categorization degree of bilingualism and language upon recall of select monolingual and bilingual. *Journal of educational psychology,* 63, 160-164.

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

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

Stocking, M.L. y Lord, F.M. (1983) Developing a common metric in Item Response Theory. *Applied psychological measurement*, *7*(2), 201-210.

Tatsuoka, M.M.(1988) *Multivariate analysis:Techniques for educational and psychological research.* New York:Macmillan

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

Aceptado el 19 de julio de 1999