*Psicothema, 1999. Vol. 11, nº 1, pp. 151-161*
*ISSN 0214 - 9915 CODEN PSOTEG*


# ADAPTACIÓN AL EUSKERA DE UNA PRUEBA DE INTELIGENCIA VERBAL

Pauli Elosúa Oliden, Alicia López Jáuregui y Esther Torres Álvarez
Universidad del País Vasco

La adaptación de pruebas psicológicas es un proceso que va más allá del estudio de la calidad lingüística de la traducción, requiere la evaluación de la equivalencia métrica entre puntuaciones y por tanto un estudio del posible sesgo. En este trabajo se estudian cada una de las fases del procedimiento seguido en la adaptación al euskera de una prueba de capacidad verbal. Ante la imposibilidad de utilización de los modelos de teoría de respuesta a los ítems en la exploración del posible sesgo, se hace uso de procedimientos alternativos de detección del funcionamiento diferencial de los ítems. Como resultados más relevantes cabría destacar el alto porcentaje de ítems con funcionamiento diferencial, la alta correlación entre los procedimientos Mantel-Haenszel y SIBTEST, la baja correlación de la regresión logística iterativa con los anteriores, y la estrecha relación entre el funcionamiento diferencial y el tipo de tarea asociado a cada uno de los ítems.

*Adaptation to basque of a verbal ability test.* The adaptation of psychological tests is a process wich goes beyond the study of the linguistic quality of translation. It is a process that requires the assessment of the metric equivalence between scores, and therefore, a study of the possible bias. In this work the adaptation to Basque of a verbal ability test is studied. The impossibility of using item respone theory models has led to the need to utilise alternative detection procedures of differential item functioning. The high percentage of ítems with differential functioning, together with the high correlation between the Mante-Haenszel statistic and Sibtest as well as the low correlation of the logistic iterative regression with previous ones, and also the close relationship between the differential functioning the kind of task associated with each item can be outlined as the most relevant results.

Los problemas derivados de la aplicación de pruebas psicopedagógicas a sujetos con un idioma dominante distinto al del grupo normalizador en el que se ha basado el proceso de estandarización es un campo de estudio de reciente implantación dentro de la psicología. Las dificultades asociadas a la administración de tests en un idioma menos familiar que el de uso habitual las encontramos referidas por primera vez en los trabajos de Sánchez (1932, 1934) relativos a las minorías hispanas en los EE.UU., si bien hasta la década de los 70 las investigaciones no prosperan en esta área y es en 1985 cuando la American Psychological Association (APA) les dedica un apartado

Correspondencia: Pauli Elosúa Oliden
Facultad de Psicología
Universidad del País Vasco
20009 San Sebastián
E-mail: pspelolp@sc.ehu.es

en sus Standards haciendo hincapié en su relevancia.

La influencia que los factores lingüísticos pueden ejercer sobre la ejecución de pruebas psicológicas hace necesaria su evaluación con el fin de poder asegurar una correcta interpretación de las puntuaciones obtenidas. Estos factores jamás han de influir negativamente sobre los resultados de una prueba a modo de variable contaminante que altere o distorsione sus características métricas; por el contrario, siempre deben maximizar las posibilidades de una correcta ejecución.

Estas consideraciones, junto con la constatación de la carencia de pruebas construidas originalmente en euskera y el consiguiente empleo generalizado de adaptaciones, han motivado el presente trabajo. En él se pretenden estudiar los aspectos relacionados con los problemas que conlleva el proceso de evaluación de la equivalencia métrica, a través de la adaptación al euskera de una prueba psicopedagógica. En este caso se considera que el idioma de aplicación es la única condición que varía entre las aplicaciones de las versiones original y adaptada. Dado que el entorno cultural de los sujetos destinatarios de pruebas psicológicas administradas dentro del proceso de escolarización es el mismo en individuos castellanoparlantes y *euskaldunes*, las posibles divergencias de resultados que puedan hallarse serán debidas a las características de cada uno de los idiomas de aplicación. Focalizamos la atención en las diferencias existentes entre un idioma con amplia tradición literaria y perfectamente normalizado como es el castellano, y otro, que si bien está considerado como una de las lenguas vivas más antiguas de Europa, posee escasa tradición literaria y un muy reciente proceso de normalización lingüística.

Para la consecución de estos objetivos se adapta al euskera un test originalmente pensado y construido en castellano, evitando de este modo el influjo de otro idioma y cultura ajenos al entorno natural de los sujetos experimentales y centrando el interés en los problemas de adaptación idiomática.

La investigación se lleva a cabo siguiendo las siguientes pautas de actuación generales:

a. Selección de una prueba que se adecue a los requisitos mencionados.

b. Adaptación de la prueba al euskera, y evaluación de la equivalencia lingüística entre ambas. Retrotraducción.

c. Selección de la muestra y obtención de los datos.

d. Estudio exploratorio del sesgo, a través de la aplicación de diversos procedimientos de detección del funcionamiento diferencial de los ítems.

e. Estudio confirmatorio del sesgo, haciendo para ello uso de los métodos de juicio que permitirán a partir de los ítems con funcionamiento diferencial, verificar la existencia o ausencia del mismo.

## Método

### *Sujetos*

Los sujetos experimentales son aquellos que cursan estudios en 4º, 5º y 6º de enseñanza primaria. Se ha seleccionado una muestra de 1.806 sujetos con edades comprendidas entre los 9 y 11 años, repartidos por el territorio de la C.A.P.V. Al territorio histórico de Gipuzkoa pertenecen 810, a Bizkaia 345 y a Araba 651.

Del total de la muestra 1.155 corresponden al grupo de bilingües *euskaldunes* y será identificado como muestra D. Los 651 restantes forman el grupo de monolingües castellanos identificados como muestra A. La denominación de las muestras (A y D) está determinada por los modelos lingüísticos que define 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.

Tras la aplicación y una vez rechazados los cuestionarios incompletos, la muestra queda reducida a 1.451 sujetos, de los que 532 constituyen la muestra A y 919 la muestra D.

## Diseño

El diseño empleado es el definido como *sujetos monolingües en castellano y euskera realizan la prueba original y adaptada* (Hambleton, 1993). En el mismo, el grupo de referencia lo forman niños cuya lengua materna es el castellano y han sido además educados en el modelo A. El grupo focal esta constituido por niños *euskaldunes*. Es preciso reparar en el hecho de que el grupo focal no está formado por monolingües en euskera, sino por bilingües, dado que la situación de contacto de lenguas actual no permite tal condición. Es sin embargo la muestra que más se ajusta al diseño empleado.

## Instrumentos

### *Prueba*

La prueba seleccionada es la *Batería de aptitudes diferenciales y generales Elemental (B.A.D.Y.G.)* (Yuste, 1988) que está compuesta por ocho subpruebas con las cuales se pretende ofrecer una cuantificación de la inteligencia general a través de dos escalas, inteligencia general verbal e inteligencia general no verbal.

De las ocho subpruebas de que consta la batería, analizamos la subescala *Habilidad Mental Verbal (H.M.V.)* , cuya versión en euskera se denomina *Hitzezko Adimen Trebetasuna (H.A.T.)*. Es de lápiz y papel y de aplicación colectiva.

La subescala habilidad mental verbal, es una prueba específica para medir la inteligencia verbal. Consta de 40 ítems de dificultad creciente, con seis alternativas de respuesta de las cuales sólo una es correcta. Los ítems se clasifican en función de las distintas tareas propuestas tal y como se muestra en la *tabla 1*.

## Recursos Informáticos

Todos los análisis se han efectuado con un PC-486DX bajo el sistema operativo MS-DOS v.6.2. y con la utilización del siguiente software:

*Tabla 1*
Media aritmética, varianza y consistencia de cada una de las tareas de H.M.V.

| | | Muestra A | | | Muestra D | | |
|---|---|---|---|---|---|---|---|
| | ítems | $\bar{X}$ | $S^2_X$ | α | $\bar{X}$ | $S^2_X$ | α |
| Constancia de una característica | 9 | 5,93 | 3,83 | 0,58 | 3,69 | 3,24 | 0,47 |
| Series lógicas con números | 7 | 5,24 | 1,80 | 0,55 | 4,99 | 1,65 | 0,54 |
| Ordenar palabras sueltas | 6 | 3,86 | 1,47 | 0,48 | 2,80 | 0,93 | 0,30 |
| Hallar género | 8 | 5,72 | 2,52 | 0,49 | 4,74 | 2,78 | 0,50 |
| Problemas numéricos | 7 | 4,07 | 3,19 | 0,65 | 2,99 | 2,16 | 0,54 |
| Problemas espacio-temporales | 3 | 2,33 | 0,69 | 0,40 | 1,92 | 0,74 | 0,29 |
| Habilidad mental verbal | 40 | 27,17 | 43,88 | 0,86 | 21,13 | 31,72 | 0,80 |

• *SPSS/PC V4.0+* paquete estadístico.

• *PC-BILOG 1.1* (Mislevy y Bock, 1986) Programa para el análisis de ítems y estimación de parámetros de habilidad bajo modelos logísticos.

• *MHDIF* (Hidalgo, 1994). Aplicación del procedimiento de Mantel-Haenszel para la detección del FDI.

• *SIBTEST* (Stout y Roussos, 1995), basado en el procedimiento del mismo nombre.

• *BIAS* (Lucassen, 1991). Aplicación del procedimiento logit iterativo para la detección del FDI.

## Adaptación de la prueba

En la adaptación de H.M.V. se han seguido las directrices generales marcadas por la retrotraducción (Brislin, 1970):

1.– La prueba original en castellano, Habilidad Mental Verbal ha sido traducida al euskera por un grupo de licenciados bilingües. La traducción ha sido fundamentalmente literal, intentando mantener el espíritu de la prueba original.

2.– Otro grupo de licenciados bilingües que no ha participado en la adaptación anterior, ha retrotraducido al castellano la versión obtenida en euskera.

3.– Se analizan las igualdades/diferencias entre ellas, con el fin de equipararlas. En la igualación de las pruebas han participado los traductores de las mismas y un profesor de enseñanza primaria. De este modo, en los casos en que un concepto o palabra podía ser traducido al euskera de varias maneras, se ha tenido en cuenta la población receptora y el criterio utilizado en la elección de la más adecuada ha sido el de idoneidad.

4.– Finalmente, un filólogo, traductor profesional, ha analizado la exactitud y coherencia lingüísticas de la prueba adaptada, corrigiendo los errores y deficiencias detectados.

## Resultados

En los análisis previos a la detección del funcionamiento diferencial de los ítems (FDI) se realizó una exploración general del comportamiento de los sujetos en las subpruebas así como de las propiedades psicométricas de las mismas, fiabilidad y estructura factorial.

Con esta simple descripción recogida en la *tabla 1*, se puede observar que la prueba H.M.V. ha resultado más difícil para los sujetos *euskaldunes* que para los del grupo A; la diferencia entre las medias es de una desviación estándar (t= 3,18; p≤0,001).

La fiabilidad de las subescalas y grupos de tareas se estima mediante el coeficiente alfa de Cronbach (α), que arroja los valores de 0,80 y 0,86 en las muestras euskaldun y castellanoparlante respectivamente. Analizando la diferencia entre los valores obtenidos para la misma prueba en las dos muestras mediante el estadístico propuesto por Feldt (1969), el valor 0,7037 nos lleva a aceptar la hipótesis nula de la igualdad entre los índices de consistencia interna.

Para comprobar la unidimensionalidad sometemos la matriz de correlaciones phi entre ítems a un análisis de componentes principales. Los resultados se muestran en la *tabla 2.*

*Tabla 2*
Varianza explicada por los tres primeros factores

| FACTORES | GRUPO DE REFERENCIA | | GRUPO FOCAL | |
|---|---|---|---|---|
| | Valores propios | % de varianza explicada | valores propios | % de varianza explicada |
| 1 | 6,6622 | 16,7 | 5,13231 | 12,8 |
| 2 | 2,3028 | 5,8 | 2,64054 | 6,6 |
| 3 | 1,58209 | 4,0 | 1,74333 | 4,4 |

En las dos muestras estudiadas se detecta la presencia de un factor principal y un punto de inflexión en la gráfica de valores

propios que se sitúa en el segundo factor, si bien la varianza explicada en la muestra A es mayor que la que explica en la muestra D. Se extraen 13 factores en la muestra monolingüe castellana que explican el 55,6% de la varianza total; en la muestra D los factores con un valor propio mayor que la unidad son 12, utilizados para dar cuenta del 50,3% de variabilidad.

Estos resultados cuestionan la existencia de unidimensionalidad. Atendiendo a los diversos índices que existen para su evaluación (Hattie, 1984, 1985) y aplicando el propuesto por Lord (1980), a saber, la razón entre la diferencia de los dos primeros valores propios y la diferencia entre el segundo y el tercero, se obtienen unos índices de 6,04 y 2,77 en las pruebas HMV y HAT. Con estos resultados, poco se puede concluir salvo, el mayor grado de unidimensionalidad de la prueba en castellano con respecto a la adaptada al euskera, pues no existe ningún test que evalúe su significancia estadística.

La comparación de las estructuras factoriales se efectúa con el coeficiente de congruencia y el índice de congruencia de Burta y Tucker (Pine, 1977; Rummel, 1970; Wrigley y Neuhaus, 1955), que adoptan los valores de 0,1283216 y 0,9521781 respectivamente. La ausencia de pruebas de significatividad nos deja sin criterios para la adopción de conclusiones estadísticas firmes. En la mayoría de los casos, se acepta que para tomar dos estructuras factoriales como equivalentes, el coeficiente de congruencia entre ellos ha de acercarse a 0 y el índice de congruencia no ha de apartarse del valor 1. En esta aproximación, sin embargo, no existe ningún punto de corte que marque la diferencia entre equivalencia y falta de ella.

Ante esta indeterminación, Schneewind y Cattell (1970), ofrecen la tabla guía de significación del índice de congruencia. Aplicándola a nuestros casos concretos, 40 ítems, tendríamos que aceptar la equivalencia de las estructuras factoriales, con una probabilidad de 0,999 .

Estas conclusiones sin embargo, no legitiman la aseveración de ausencia de sesgo en la adaptación. Al estudio del coeficiente de fiabilidad y estructura factorial es necesario agregar una evaluación pormenorizada de cada uno de los ítems que componen la prueba.

## Detección del funcionamiento diferencial de los ítems

Dado que el marco teórico más adecuado para el estudio del funcionamiento diferencial de los ítems lo proporciona la teoría de respuesta al ítem, evaluamos la adecuación entre varios de estos modelos y los datos. Los modelos seleccionados son los logísticos de uno y dos parámetros. No aplicamos el modelo logístico de tres parámetros dado que en ítems con seis alternativas de respuesta el peso del azar es mínimo y además la estimación del parámetro de pseudo-azar que daría cuenta del mismo es todavía hoy muy débil (Muñiz, 1990; Kolen, 1981; Thissen y Wainer, 1982). El procedimiento de estimación utilizado ha sido la estimación marginal de máxima verosimilitud (Bock y Aitkin, 1981) implementado en el programa BILOG 1.1 (Mislevy y Bock, 1986).

Aunque en la predefinición del tamaño de la muestra a la que se ha administrado la prueba adaptada al euskera se ha pretendido asegurar la competencia de los procedimientos basados en la T.R.I., no existe correspondencia entre el modelo y los datos.

El modelo de un sólo parámetro con $\chi^2_{\text{muestra A}}= 491,7$ y $\chi^2_{\text{muestra D}}= 674,8$ no se adecúa a ninguna de las muestras, mientras que en el modelo de dos parámetros el valor $\chi^2= 256,5$ presenta una probabilidad de adecuación en la muestra castellanoparlante monolingüe $p\leq0,6829$. En la muestra euskaldun, aún siendo el doble en tamaño, no existe tal ($\chi^2_{\text{muestra D}}= 397,2$ $p\leq0,0001$). Esto imposibilita la utilización de cualquier procedi-

miento derivado de la TRI, y nos obliga a hacer uso de técnicas alternativas. Entre los procedimientos existentes hemos seleccionado el estadístico Mantel-Haenszel (Holland y Thayer, 1988), logit-iterativo (Van der Flier, Mellenbergh, Adèr y Wijn, 1984), la regresión logística (Swaminathan y Rogers, 1990) iterativa y SIBTEST (Shealy y Stout, 1993). Los criterios utilizados en su elección han sido la tasa de uso, la efectividad y la disponibilidad. De este modo hemos desestimado los basados en el análisis del chi-cuadrado (Scheuneman, 1979) y los derivados de la teoría de contaminadores (Oort, 1992).

• El estadístico Mantel-Haenszel renovado ha sido aplicado mediante el programa MHDIF (Fidalgo, 1994b). Este programa incluye, además del proceso de purificación de la puntuación (Holland y Thayer, 1988), las aportaciones propuestas por Mazor, Clauser y Hambleton (1994) para la detección tanto del funcionamiento diferencial uniforme como del no uniforme.

Cuando el índice de funcionamiento diferencial $\alpha_{MH}$ es mayor que la unidad, los odds son mayores para el grupo de referencia, muestra A, que para el grupo focal, muestra D; mientras que un valor por debajo de uno va unido a los ítems con FD de signo positivo en favor de la muestra euskaldun.

• El logit iterativo se ha aplicado con el programa BIAS V1.0 (Lucassen, 1991). El programa BIAS, además de purificar las puntuaciones de modo iterativo (Van der Flier, Mellenbergh, Adèr y Wijn, 1984), posibilita el truncamiento de las distribuciones de los grupos de referencia y focal cuando son muy diferentes, haciendo de este modo más efectiva la detección del funcionamiento diferencial del ítem.

Una vez dividida la puntuación total en ocho intervalos, se procede a aplicar un procedimiento de detección iterativo que converge en la iteración 22 con un número de ítems con FD detectados de 21.

• En el caso de la regresión logística, los cálculos se han realizado mediante el SPSS/PC V.4.0. En este procedimiento, para calcular el criterio interno, es decir, la puntuación total, no se tiene en cuenta el ítem que se analiza. Si bien es cierto que el proceso de purificación de la puntuación ha sido ya utilizado, en el trabajo en el que ha sido evaluado, la eliminación de ítems contaminantes se ha llevado a cabo en función del valor obtenido por el estadístico de Wald (Gomez y Navas, 1995), en este trabajo sin embargo utilizamos como criterio el valor del chi-cuadrado residual. La purificación de la puntuación se ha llevado a cabo siguiendo los pasos:

1. Una vez aplicado el modelo sin FDI a todos los ítems, se estudia su grado de adecuación a través del chi-cuadrado residual, detectándose el ítem más aberrante.
2. Se elimina ese ítem del cálculo de la puntuación total y se aplica nuevamente el modelo anterior a todos los ítems.
3. En función de los valores obtenidos, volvemos al paso 2, hasta que en dos iteraciones consecutivas obtengamos el mismo resultado o hasta eliminar del cálculo todos los ítems con funcionamiento diferencial.

En la aplicación de este procedimiento han sido necesarias 29 iteraciones, en cada una de las cuales el ítem objeto de estudio ha sido excluido del cálculo de la puntuación total. En la iteración número 30 el número de ítems con FD ha sido de 29.

• El procedimiento SIBTEST ha sido aplicado mediante el programa informático del mismo nombre (Stout y Roussos, 1995). Guiados por el principio de efectividad, también en este caso hemos utilizado un procedimiento de cálculo iterativo que pretende mejorar la eficacia en la detección del FDI:

1. Aplicar el SIBTEST a todos los ítems, calculando la puntuación total en cada uno

de los casos excluyendo de la misma el ítem objeto de estudio.

Como resultado de esta primera aplicación, se consiguen dos subpruebas; la subprueba válida, formada por los ítems que no tienen funcionamiento diferencial, y *la prueba objeto de estudio*, formada por todos los ítems que tienen FD.

2. Uno a uno, se analizan los ítems que componen la prueba objeto de estudio, tomando como variable condicionante la subprueba válida.

3. Se analizan solamente los ítems que forman la subprueba válida, calculando la puntuación total con los n-1 ítems que la forman una vez excluido el ítem objeto de estudio.

4. Con los resultados obtenidos en los dos pasos anteriores, se reconstruyen la segunda subprueba válida y la segunda subprueba objeto de estudio.

5. Los dos últimos pasos se repiten sin término hasta que en dos iteraciones consecutivas se obtenga el mismo resultado.

El proceso de purificación converge en siete iteraciones. El número de ítems detectados por este procedimiento ha sido de 23. El índice $b_0$ con valores negativos indica funcionamiento diferencial a favor del grupo focal o muestra euskaldun, mientras que el índice positivo significa que la diferencia de medias o proporciones de respuesta correcta al ítem a través de todos los niveles de puntuación favorece al grupo de referencia o prueba castellana.

La tabla 3 resume los resultados obtenidos en los cálculos efectuados con un nivel de riesgo de 0,05. En negrilla aparecen los ítems catalogados con funcionamiento diferencial, y en la última fila se muestra el porcentaje de funcionamiento diferencial que detecta cada procedimiento (P.D.).

En una observación preliminar puede advertirse por un lado el notable porcentaje de FDI detectado por todos los procedimientos empleados, y por otro, la falta de unanimidad en la catalogación de los ítems con FD. Una de las causas de tal desacuerdo puede hallarse en la influencia que el porcentaje de ítems con FD ejerce sobre la eficacia de ca-

*Tabla 3*
Funcionamiento diferencial según los distintos procedimientos

| Ítem | Mantel-Haenszel $a_{MH}$ | Logit-iterativo $L^2$ | Reg.logis. $c^2_{residual}$ | Sibtest $b_0$ |
|---|---|---|---|---|
| 1 | 1.10 | **17.347** | **16.587** | 0.006 |
| 2 | 1.05 | 5.829 | 7.603 | 0.026 |
| 3 | **0.42** | 14.726 | 4.105 | **-0.047** |
| 4 | 1.15 | 6.810 | **11.642** | 0.028 |
| 5 | **0.48** | 13.405 | 0.811 | **-0.083** |
| 6 | 1.11 | 7.447 | **19.042** | 0.040 |
| 7 | 1.06 | 7.203 | 4.929 | 0.010 |
| 8 | **0.54** | **21.379** | 1.910 | **-0.041** |
| 9 | **0.49** | **25.481** | 8.008 | **-0.058** |
| 10 | **1.62** | **26.889** | **76.236** | **0.112** |
| 11 | **0.52** | **16.667** | 1.175 | **-0.074** |
| 12 | 1.22 | **15.789** | **26.192** | **0.062** |
| 13 | 1.07 | 12.440 | **10.390** | 0.021 |
| 14 | **0.28** | **54.471** | **14.378** | **-0.152** |
| 15 | 0.98 | **16.031** | **13.647** | -0.005 |
| 16 | **2.89** | **95.775** | **182.215** | **0.240** |
| 17 | 1.27 | **25.648** | **34.750** | **0.092** |
| 18 | **0.35** | **40.447** | **7.835** | **-0.121** |
| 19 | **1.28** | **16.138** | **26.698** | **0.059** |
| 20 | **2.38** | **72.085** | **134.564** | **0.215** |
| 21 | **0.41** | **48.675** | **6.094** | **-0.135** |
| 22 | 0.98 | 10.211 | **21.497** | 0.038 |
| 23 | **0.37** | **33.443** | 6.670 | **-0.193** |
| 24 | 1.09 | 11.718 | **31.710** | 0.044 |
| 25 | **0.57** | **20.945** | 1.091 | **-0.125** |
| 26 | **0.39** | **41.863** | **13.508** | **-0.187** |
| 27 | 1.08 | **23.037** | **27.965** | -0.008 |
| 28 | **0.38** | **33.927** | 0.638 | **-0.199** |
| 29 | **1.42** | **34.810** | **77.465** | 0.031 |
| 30 | **2.61** | **73.249** | **154.750** | **0.159** |
| 31 | 0.83 | **16.110** | **19.649** | -0.026 |
| 32 | 0.85 | 11.004 | **10.294** | -0.048 |
| 33 | **1.42** | **30.776** | **41.577** | 0.013 |
| 34 | **4.50** | **102.188** | **134.451** | **0.131** |
| 35 | 1.10 | 13.426 | **18.921** | -0.008 |
| 36 | 1.16 | 6.719 | **14.288** | -0.018 |
| 37 | **0.47** | 15.209 | 5.744 | **-0.074** |
| 38 | 0.81 | 14.900 | **26.529** | **-0.058** |
| 39 | **0.81** | 13.819 | **25.719** | **-0.057** |
| 40 | 0.73 | 8.371 | **14.972** | -0.020 |
| % | 55 | 60 | 72 | 55.5 |

da una de las técnicas. Influencia que se refleja principalmente en el incremento de los errores tipo I o falsas detecciones (Mazor, Clauser y Hambleton, 1994). Además encontramos en nuestro datos otra característica que repercute negativamente en el grado de operatividad, la diferencia significativa entre las distribuciones .

Un estudio más profundo de la efectividad y grado de concordancia entre todos los procedimientos nos lleva a estudiar las relaciones entre ellos. Con este fin se analizan las correlaciones entre los resultados para evaluar posibles divergencias o convergencias . Estas se muestran en la *tabla 4*.

*Tabla 4*
Correlación entre procedimientos

| | Estad. MH. | Logit iter. | Regre.log. |
|---|---|---|---|
| Estad. MH. | | | |
| Logit iter. | 0.4924 | | |
| Regre.log. | -0.3320 | 0.0686 | |
| Sibtest | 0.7472 | 0.4336 | -0.3030 |

Los valores que aparecen en la matriz de correlación son significativos en los casos del estadístico MH, logit iterativo y SIBTEST. Sin embargo, las correlaciones de la regresión logística con los demás procedimientos son muy bajas. Dada la proximidad entre procedimientos, esperábamos obtener una correlación mas estrecha entre el logit iterativo y la regresión logística. Al interpretar la correlación de 0,0686 obtenida es preciso tener en cuenta que para evitar la influencia de la diferencia de las distribuciones originales, el programa BIAS en el que está implementado el logit iterativo, posibilita el truncamiento de las distribuciones a fin de mejorar las tasas de detección. En la aplicación de la regresión logística no se han efectuado truncamientos de este tipo, y aparece negativamente influenciada por la diferencia entre las distribuciones.

Atendiendo al nivel de concordancia alcanzado entre los otros tres procedimientos MH, logit-iterativo y SIBTEST, aún en el peor de los casos se sitúa en torno al 75%. expondremos las posibles causas de sesgo en función de los resultados aportados por los mismos.

## Discusión

Analizaremos por tanto 29 ítems que parecen mostrar funcionamiento diferencial, el 72,5% de los ítems de la prueba. Un porcentaje desmesurado. De ellos, y es lo que realmente no esperábamos, 17 presentan un funcionamiento diferencial favorable a la prueba adaptada y 12 propician a la muestra castellanoparlante.

Dado que los ítems de la prueba de H.M.V. están clasificados en diferentes grupos en función a la la tarea que exigen a los sujetos, nos parece interesante examinar si existe alguna relación entre ésta y el FDI. En la tabla 5, además de los ítems que forman cada tarea, aparecen en negrilla los ítems con FD acompañados de un signo; cuando es + el sentido del funcionamiento diferencial favorece a la muestra euskaldun,

*Tabla 5*
Sentido del funcionamiento diferencial según el tipo de tarea

| | ítems |
|---|---|
| Constancia de una característica | 6 **-10 -16 -17 -19 -20 -27 -33** 35 |
| Series lógicas con números | **+3 +9 +14 +18 +26 +28** 40 |
| Ordenar palabras sueltas, formando una frase correcta | **-1** 4 7 **-30 -34 +37** |
| Hallar el género o criterio de clasificación de una serie | 2 **+8 +11 -12** 22 24 **+25 +30** |
| Problemas numéricos de comprensión lógico-numérica | **+5 +15 +23 +31** 32 36 **+38** |
| Problemas de resolución espacio-temporal | 13 **+21 -29** |
| TOTAL | +17 -12 |

y cuando es - favorece al grupo de referencia o muestra castellanoparlante.

El valor de $\chi^2$= 23,003 obtenido para verificar la relación entre las variables categóricas que se recogen en esta tabla, es significativo ($p \leq 0,01$). Podemos concluir por lo tanto que existe relación entre el tipo de ítem, es decir, la clase a la que pertenece, por un lado, y la existencia y el signo del mismo por el otro. El coeficiente de contingencia es de 0,604.

Con el fin de apreciar mejor la relación, hemos clasificado los ítems guiados por un criterio más general. En esta redistribución los grupos son mas amplios; problemas o operaciones ligados a números, ítems totalmente verbales y los espacio-temporales (tabla 6).

*Tabla 6*
Sesgo y su relación con el tipo de ítem

| | Número de ítems | Sesgo positivo | Sesgo negativo |
|---|---|---|---|
| Problemas numéricos | 14 | 11 | 0 |
| Problemas verbales | 23 | 5 | 11 |
| Problemas espacio-temporales | 3 | 1 | 1 |
| Total | 40 | 17 | 12 |

De este modo parece patente la estrecha relación entre el tipo de tarea y la existencia de funcionamiento diferencial. Relación que obliga a un análisis de contenido y forma de cada uno de los ítems que componen la escala, de modo que arroje luz y criterios específicos para la correcta redacción de los ítems en euskera.

Tras este trabajo podemos afirmar que en el proceso de adaptación al euskera de pruebas psicopedagógicas, la evaluación correcta y estricta de su calidad lingüística no garantiza por sí misma la equivalencia métrica entre versiones. Para asegurar la equivalencia entre puntuaciones, es imprescindible, dentro del examen de la validez de las pruebas, el análisis del sesgo, a través de los instrumentos que nos habilitan los procedimientos estadísticos condicionales para la detección del funcionamiento diferencial de los ítems, así como de los procedimientos de juicio que puedan dar una explicación del mismo.

Esto quiere decir, que los procedimientos habitualmente utilizados para analizar la igualdad entre pruebas, a saber, el análisis factorial y el estudio de la fiabilidad, no son de ningún modo suficientes en la conclusión de la existencia de equivalencia métrica. Si bien es cierto que la igualdad es condición necesaria. Si el sesgo se da en el nivel de los ítems, estos procedimientos que consideran la prueba en su totalidad, no lo detectarán, orientándonos hacia una equivalencia métrica que en realidad no existe.

Es evidente que toda adaptación comienza con una traducción lingüística. La prueba obtenida, no es otra cosa que *la prueba piloto* de la que será la prueba definitiva. Este es un hecho que hay que reconocer en el proceso de análisis de la equivalencia métrica que pretendemos obtener. Una vez administrada la prueba piloto, es imprescindible dentro del análisis de la validez, investigar *el posible sesgo*. Para esta investigación, podemos hacer uso del fundamento teórico y de las herramientas que nos ofrecen los modelos de teoría de respuesta al ítem, u otros procedimientos referidos en el trabajo, para, en el ámbito de lo que hemos denominado *estudio exploratorio del sesgo*, detectar el funcionamiento diferencial de los ítems.

Una vez detectado el FDI, continuamos con *el análisis confirmatorio* del sesgo. Mediante procedimientos de juicio y profundizando en el contenido de los ítems, se intentará buscar las diversas causas que hayan podido originarlo. En este nivel de análisis es aconsejable solicitar la colaboración de expertos en el área de contenido general de la prueba.

El proceso de adaptación sin embargo, no se da por concluido con el análisis del

sesgo. Es necesario continuar con la equiparación. Formando la *prueba de anclaje* con los ítems sin sesgo se colocan las puntuaciones de las dos versiones de la prueba sobre la misma escala.

Desde un punto de vista práctico, el costo de una adaptación correcta, tanto en tiempo como en recursos económicos, se convierte en un factor a considerar. En este trabajo, se ha llevado a cabo la primera fase, la adaptación lingüística de la prueba original, el análisis del funcionamiento diferencial de los ítems y la confirmación del sesgo. Los resultados obtenidos, nos obligan a dejar para un desarrollo posterior, tanto la eliminación del sesgo detectado como la equiparación de las puntuaciones. En esta etapa, se confirmarán o desecharán las hipótesis formuladas con respecto a los ítems sesgados después de la aplicación de procedimientos de juicio explicativos de los mismos.

Como conclusión básica, afirmaremos que la traducción lingüística correcta, no es suficiente garantía para la equivalencia métrica. Si se quiere confirmar la igualdad de significado de las puntuaciones obtenidas mediante las versiones original y adaptada, es necesario el análisis profundo del sesgo, donde son imprescindibles la utilización de procedimientos empíricos y de juicio.

## Agradecimientos

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

## Referencias

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

Bock, R.D. & Aitkin, M. (1981). Marginal maximuum likelihood estimation of ítem parameters. *Psychometrika, 46(4),* 443-459.

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

Elosúa, P., López, A. & Artamendi, J.A. (1994). Elebiduntasunari buruzko testaren bidez lorturiko datoen azterketa kuantitatiboa. *Tantak, 12,* 197-218.

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.

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

Gómez, J. y Navas, M.J. (1995, Abril). Detección del sesgo mediante regresión logística: purificación paso a paso de la habilidad. *Comunicación presentada al IV Symposium de Metodología de las Ciencias del Comportamiento,* Murcia, España.

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

Hammers, B. & Blanc, M. (1983). *Bilingualité et Bilinguisme.* Bruxelles: Pierre Mardaga Ed.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research,* 19, 49-78.

Hattie, J. (1985) Methodology review: Assessing unidimensionality of test and ítems. *Applied Psychological Measurement, 9(2),* 139-164.

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

Kolen, M.J. (1981). Comparation of traditional and ítem response theory methods for equa-

ting tests. *Journal of Educational Measurement, 18,* 1-11.

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

Mazor, K.M., Clauser, P.E. & Hambleton, R.K. (1994). Identification of nonuniform Differential Ítem Functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54,* 284-291.

Mislevy, R.J. & Bock, R.D. (1986). *Bilog: Ítem 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.

Oort, F.J. (1992). Using restricted factor analysis to detect ítem bias. *Methodika, VI,* 150-166.

Pine, S.M. (1977). Applications of Ítem Response Theory to problem of test bias. In D.J. Weiss(Ed.), *Applications of computerized adaptive testing* (pp.37-43); (Research Report N° 77-1). Minneapolis: University of Minnesota.

Rummerl, J.R. (1970). *Applied Factor Analysis.* Evanston, IL: Northwestern University Press.

Sánchez, G.I. (1932). Scores of Spanish-speaking children on repeated tests. *Journal of Genetic Psychology,* 40(1).

Sánchez, G.I. (1934). Bilingualism and mental measures: A world of caution. *Journal of Applied Psychology, 18.*

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

Schneewind, K & Cattell, R.B. (1970). Zum Problem der Faktoridentifikation: Verteilungen und Vertrauensintervalle von Kongruentzkoeffizienten. *Psychology Beiträge, 12,* 214-226.

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

Stout, W. & Roussos, L. (1995). SIBTEST [Computer program]. Urbana-Champaign, IL: University of Illinois, Statistical laboratory for Educational and Psychological Measurement.

Swaminathan, H. & Rogers, H.J. (1990). Detecting differentail ítem functioning using logistic regresion procedures. *Journal of Educational Measurement, 27(4),* 361-370.

Thiseen, D. & Wainer, H. (1982). Some standards errors in Ítem Response Theory. *Psychometrika, 47(4),* 397-412.

Van der Flier, H., Mellenbergh, G.J., Adèr, H.J. & Wijn, M. (1984). An iterative ítem bias detection method. *Journal of Educational Measurement, 21(2),* 131.145.

Wrigley, C. & Neuhhaus, J.O. (1955). The matching of two sets of factors. *Contract Memorandum Report.* University of Illinois.

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

*Aceptado el 15 de junio de 1998*