Psicothema
2000. Vol. 12, Supl. nº 2, pp. 214-219

ISSN 0214 - 9915 CODEN PSOTEG

Comparación del estadístico Mantel-Haenszel y la regresión logística en el funcionamiento diferencial de los ítems en dos pruebas de aptitud intelectual en un contexto bilingüe

Doris Ferreres Traver, Vicente González Romá y Juana Gómez Benito*
Universidad de Valencia y * Universidad de Barcelona

Uno de los temas relevantes en la práctica psicométrica es el estudio de las propiedades psicométricas de los tests aplicados a poblaciones con características diferentes a aquéllas para las que éstos fueron creados, prestando un especial interés a los ítems con problemas de funcionamiento diferencial (DIF). Nuestro estudio pretende abordar esta problemática en una situación de lenguas en contacto, como es la Comunidad Valenciana. En ella, la evaluación de las aptitudes intelectuales se realiza mediante la aplicación de tests psicológicos redactados y baremados en castellano a alumnos cuya lengua familiar y/o escolar es el Valenciano. El presente trabajo pretende analizar la congruencia del estadístico Mantel-Haenszel y la regresión logística en la detección del DIF en dos pruebas elaboradas y baremadas en castellano y de uso habitual en el ámbito educativo valenciano. Los resultados serán interpretados atendiendo a las características de los ítems con DIF, y a los resultados conseguidos por distintos estudios de simulación. A su vez, se aportarán recomendaciones a considerar en futuros estudios de pruebas psicológicas con problemas de DIF en sus ítems.

Mantel-Haenszel statistic and the logistic regression in the detection of DIF in two aptitude tests. Psychometricians have stressed the need for studying the psychometric characteristics of items applied to groups of subjects with characteristics different to those of the original target population, focusing on the detection of items which function differentially across groups (DIF). This question was studied in a situation of languages in contact, such as the Valencian Community (Spain). In this region, it is quite frequent to apply ability tests developed in Spanish language to children whose family and/or school language is Valencian. This paper analyzes the congruence of the Mantel-Haenszel statistic and the logistic regression in the detection of DIF in two scales, written and validated in Spanish, and frequently applied in the Valencian educational context. Results were interpreted attending to the specific characteristics of DIF items and simulation studies. Several recomendations to consider in future investigations of psychological scales with DIF items are given.

En la actualidad, la aplicación de instrumentos de medida originariamente elaborados para grupos con una lengua y/o cultura mayoritaria a ciertos grupos minoritarios con una lengua y/o cultura diferente es una práctica muy extendida en el ámbito de la Psicología y la Educación, y siempre bajo el supuesto del funcionamiento equivalente de los ítems del test administrado a través de los grupos. Sin embargo, este supuesto asumido y aceptado por una gran mayoría no siempre ha sido confirmado, descubriéndose que algunos tests o parte de sus ítems estaban sesgados hacia determinados grupos de la población. Esta circunstancia propició que uno de los temas más relevantes en la práctica psicométrica fuera el estudio de las propiedades psicométricas de los tests aplicados a poblaciones con características diferentes a aquéllas para las que éstos fueron creados, prestando un especial interés a la identificación de los ítems con problemas de funcionamiento diferencial (*Differential Item Functioning*, DIF, a partir de ahora). Una revisión de la literatura sobre este tema y del área de bilingüismo (Botempo, 1993; Budgell, Raju y Quartetti, 1995; Candell y Hulin, 1987; Ellis, 1989; Hulin y Mayer, 1986) nos demuestra que los contextos educativos donde coexisten más de una lengua (escenarios conocidos como situaciones de contacto de lenguas) son uno de los ámbitos más habituales de estudio del DIF. La enorme trascendencia de este fenómeno se pone de manifiesto cuando se constata la excesiva frecuencia con que un instrumento de medida desarrollado con normas monolingües es aplicado a grupos mayoritarios y a otros más o menos minoritarios, que conviven en una misma zona geográfica pero que poseen una lengua familiar y/o de escolarización diferente.

Nuestro estudio pretende abordar esta problemática en una realidad empírica con las características propias de una situación de contacto de lenguas, como es la Comunidad Valenciana. Ésta se caracteriza por ser una comunidad bilingüe donde el uso y aprendizaje de sus dos lenguas oficiales, el castellano y el valenciano, es

Correspondencia: Doris Ferreres Traver
Facultad de Psicología
Universidad de Valencia
46010 Valencia (Spain)
E-mail: doris@uv.es

un hecho bastante común. Sin embargo, su presencia difiere notablemente según la zona geográfica considerada, circunstancia determinante en el uso y empleo del valenciano como lengua de comunicación y, no menos importante, como lengua de escolarización en los modelos educativos implantados en la Comunidad Valenciana. En esta comunidad, el uso de los tests psicológicos para la evaluación de las aptitudes intelectuales es un hecho bastante generalizado en la mayoría de los centros escolares valencianos. Sin embargo, esta práctica educativa desarrollada en un contexto bilingüe presenta una importante peculiaridad: *la aplicación de tests elaborados y baremados en castellano a alumnos en proceso de escolarización de programas de enseñanza bilingue.* Las repercusiones educativas y sociales que se derivan del uso inadecuado de los tests psicológicos resultan bastante evidentes. Así pues, si ciertos grupos son perjudicados por sus características lingüísticas cuando sus aptitudes intelectuales son medidas debido a que los tests empleados contienen ítems con funcionamiento diferencial (DIF), es probable que las decisiones que se tomen a partir de estas medidas no sean adecuadas y, en consecuencia, se tienda a favorecer o subestimar injustamente las competencias de los sujetos pertenecientes a un grupo lingüístico concreto en relación al otro.

Métodos y técnicas estadísticas en la detección del DIF

En la últimas décadas, el estudio del DIF, y por extensión del DFT de los tests, ha recibido una amplia atención por parte de los profesionales de la medida. Para una revisión metodológica de los principales métodos estadísticos del DIF puede consultarse los textos de Fidalgo (1996), Gómez e Hidalgo (1997), y Millsap y Everson (1993). Tomando la revisión realizada por Millsap y Everson (1993), podemos apreciar que la mayoría de las contribuciones producidas en este campo pueden clasificar en 2 categorías: *a) Métodos de Invarianza Condicional Observada (ICO)*, que evalúan el DIF con un modelo de medida que relaciona las puntuaciones en el ítem con la variable latente que pretende medir el test, y *b) Métodos de Invarianza Condicional No observada (ICN)* que evalúan el DIF sin necesidad de especificar un modelo de medida que relacione la variable medida por el test con la puntuación obtenida en el ítem. Dentro estos métodos, también denominados análisis de tablas de contingencia (TC), se incluyen entre otros el estadístico Mantel-Haenszel (Holland y Thayer, 1986) y la regresión logística (Rogers y Swaminathan, 1993; Swaminathan y Rogers, 1990).

Ante tal cantidad de métodos y procedimientos estadísticos, resultaría interesante conocer qué métodos de análisis son los más eficaces y bajo qué condiciones en la detección y evaluación del DIF. Recordemos que nuestro estudio pretende abordar el problema del DIF desde un punto de vista aplicado, y al trabajar con datos reales es imposible conocer la relación exacta entre el número de detecciones correctas y los errores de Tipo I (falsos positivos (FP)) y de Tipo II (falsos negativos (FN)) existentes. Por ello, no se puede conocer la capacidad de detección de cada procedimiento estadístico, pero sí es posible una valoración basada en el grado de concordancia mostrado por los distintos métodos de análisis de DIF utilizados. Por otra parte, los estudios empíricos en su gran mayoría no suelen cumplir los requisitos muestrales y técnicos necesarios para aplicar adecuadamente las técnicas basadas en la TRI, lo que provoca que los análisis de tablas de contingencia (TC) se presentan como una de las alternativas más asequibles. En esta categoría, el estadístico MH y la regresión logística son una de las opciones más firmes en la detección y evaluación del DIF.

Así pues, dentro de esta línea de investigación se inscribe el presente estudio en el que se analiza el grado de congruencia existente entre el estadístico Mantel-Haenszel y la regresión logística en la detección del DIF en dos pruebas de aptitud intelectual, elaboradas y baremadas en castellano y de uso habitual en el ámbito educativo valenciano.

Método

Instrumento de medida

La prueba utilizada fue la *Batería de Aptitudes Diferenciales y Generales (BADYG).* Su selección residió en el elevado uso de la misma en los centros escolares y Gabinetes Psicopedagógicos consultados a través de una encuesta telefónica. Dicha batería elaborada y baremada en castellano es una de las pruebas psicológicas más utilizadas en la Comunidad Valenciana. La totalidad de las pruebas, 8 en total, presentan seis niveles diferentes de aplicabilidad escolar. Todas son pruebas de lápiz y papel y de aplicación colectiva. Sin embargo, nuestro estudio centró su interés en dos de ellas: a) escala *Habilidad Mental Verbal (HMV),* compuesta por 40 ítems con un fuerte contenido verbal, y b) escala *Habilidad Mental No-Verbal (HMNV),* formada por 40 ítems gráfico-geométricos. Para ambas escalas se consideraron los niveles de aplicabilidad: Elemental y Medio.

Muestra

La selección de las escuelas participantes en este estudio se realizó atendiendo a las diferencias cualitativas y cuantitativas que existen en el uso y dominio de la lengua propia (valenciano) en la citada comunidad. Así, los distintos grupos objeto de estudio fueron definidos atendiendo a: 1) la lengua familiar (LF), y 2) la lengua de escolarización (LE). La primera variable (LF) adquiere una relevancia considerable en la Comunidad Valenciana, ya que su situación lingüística queda claramente definida por una relación de desigualdad entre la lengua castellana y valenciana. La operacionalización de esta variable se realizó mediante la aplicación de una encuesta socio-lingüística a cada escolar, donde se incluían algunas cuestiones acerca del uso de las dos lenguas oficiales a nivel familiar, escolar, coloquial, etc. Su operacionalización se realizó como sigue: a) *lengua familiar castellana (LF C):* se seleccionaron los escolares cuya lengua familiar era siempre el castellano, y b) *lengua familiar valenciana (LF V):* se tomaron aquellos alumnos cuya lengua familiar era el valenciano o al menos existía una predominancia del valenciano sobre el castellano. La segunda variable considerada fue el modelo educativo o lengua de escolarización (LE). Los programas educativos implantados en la Comunidad Valenciana pueden clasificarse en dos categorías: a) *Educación monolingüe en castellano (LE C),* cuya enseñanza es íntegramente en castellano, y b) Educación bilingüe. Dentro de ésta última, se distingue entre: 1. *Modelo de mantenimiento (LE CV),* cuya lengua de escolarización es el castellano en toda la enseñanza obligatoria, a excepción de algunas variantes que se introducen a partir de 3º de Primaria y 1º de E.S.O., donde una parte de las áreas curriculares se imparten en valenciano, y 2. *Modelo de enriquecimiento (LE V*), cuya característica principal es la adquisición y aprendizaje de las habilidades lectoescritoras en valenciano.

La totalidad de la muestra fue de 2128 alumnos escolarizados en 4º, 5º y 6º de Primaria, y 1º y 2º de Secundaria durante el curso académico 1997-1998. Respecto a la Educación Primaria, se obtuvieron datos de alumnos escolarizados en los tres modelos educativos: modelo monolingüe en castellano (N=333), modelo de mantenimiento (N=193) y modelo de enriquecimiento (N=249). En relación al nivel de Educación Secundaria, se recogieron datos de alumnos escolarizados únicamente en los modelos monolingüe (N=316) y de mantenimiento (N=285).

Así pues, se comparó el rendimiento ofrecido por grupos de alumnos cuya lengua familiar y escolar era el castellano (grupo de referencia: muestra LF C LE C) frente a: 1) escolares cuya lengua escolar era el castellano, pero su lengua familiar era el valenciano (grupo focal 1: muestra LF V LE CV), y 2) alumnos cuya lengua familiar y de escolarización era el valenciano (grupo focal 2: muestra LF V LE V). Dichas comparaciones se realizaron considerando: a) *nivel de aplicabilidad* del instrumento de medida utilizado: BADYG Elemental, aplicable a escolares de 4º, 5º y 6º de Primaria, y BADYG Medio, aplicable a alumnos de 1º y 2º de Secundaria; y b) *tipo de prueba*: una prueba compuesta por ítems de contenido verbal, denominada escala de Habilidad Mental Verbal (HMV), y una prueba formada por ítems gráficos, denominada escala de Habilidad Mental No Verbal (HMNV). La importancia de la primera variable -nivel de aplicabilidad- reside en que los alumnos en proceso de escolarización en programas bilingües están menos expuestos a la presencia y aprendizaje de la lengua castellana, y posiblemente su nivel de competencia lingüística en dicha lengua no esté suficientemente desarrollada y consolidada hasta llegar a los últimos cursos de la educación obligatoria, lo cual lleva a esperar un mayor porcentaje de ítems con DIF en las pruebas aplicadas en Educación Primaria (BADYG Elemental) que en las pruebas correspondientes a Educación Secundaria (BADYG Medio). Por su parte, la variable -tipo de prueba- adquiere una gran relevancia en el contexto analizado (situación de lenguas en contacto), ya que se espera que aparezcan items con DIF sólo en la prueba de contenido verbal.

Detección del DIF

En cuanto al análisis del DIF de las escalas utilizadas, el presente estudio optó por una detección del DIF en datos dicotómicos aplicando dos métodos de análisis: el estadístico MH y la regresión logística (RL). Respecto al estadístico Mantel-Haenszel (Holland y Thayer, 1986), la gran cantidad de literatura en torno a este método habla por sí sola. Es una de las técnicas estadísticas más utilizadas en la detección del DIF debido a la simplicidad de su cálculo e interpretación, y sus buenos resultados en tamaños muestrales pequeños (200 sujetos por grupo), además de ofrecer una cuantificación de la magnitud del DIF (cociente de razones común), y un test de significación estadística. En cambio, una de las mayores críticas a este método ha sido su incapacidad para detectar el DIF no uniforme (Swaminathan y Rogers, 1990; Rogers y Swaminathan, 1993). Una forma de erradicarlo ha sido la modificación de cálculo en el estadístico MH propuesta por Mazor, Clauser y Hambleton (1994) para la detección del DIF no uniforme. La aplicación de tales técnicas se realizó mediante el programa MH-DIF elaborado por Fidalgo (1994). Por su parte, Swaminathan y Rogers (1990) propusieron la regresión logística como método de análisis del DIF uniforme y no uniforme. En esta aproximación, el modelo permite predecir la probabilidad de una respuesta correcta a un determinado ítem en función del nivel de habilidad del examinado (θ: la puntuación total en el test), su grupo de pertenencia (G), y el término θxG, que representa la interacción entre el nivel de habilidad del examinado y su grupo de pertenencia, siendo este último término el que en caso de resultar significativo indicaría la presencia de DIF no uniforme.

Para ambos métodos, se realizó una depuración de la medida de habilidad aplicando una versión bietápica del criterio, con el fin de maximizar la probabilidad de identificar sólo aquellos ítems que realmente poseen DIF. En el presente estudio, los ítems analizados presentaban DIF cuando los estadísticos correspondientes (ji-cuadrado MH y Wald) eran estadísticamente significativos, a un nivel de significación de 0.05/nº ítems de la escala (HMV: p= 0.05/40= 0.0025; HMNV: p= 0.05/40= 0.0025). Pero, debido a que los programas empleados trabajan únicamente con los niveles 0.005 y 0.001, este estudio aplicó el nivel de significación de 0.005. La justificación de este proceder es un intento de prevenir la aparición del error de Tipo I y, por tanto, asegurar que los ítems presentan DIF cuando realmente existe.

Resultados

Análisis descriptivos, Fiabilidad y Unidimensionalidad

Estudios previos (Ferreres, 1998; Ferreres, González, Lloret y Gómez, 1999) confirman la homogeneidad de las muestras en cuanto a la variable género y edad. Los análisis descriptivos realizados mostraron que la escala HMNV perteneciente al BADYG Elemental ofrecía un mejor rendimiento en los dos grupos focales (LF V LE CV y LF V LE V) que en el grupo de referencia (LF C LE C), existiendo diferencias estadísticamente significativas (LF C LE C vs. LF V LE CV $t=-2.78$, $p\leq0.006$; LF C LE C vs. LF V LE V $t=-2.28$, $p\leq0.023$). Por lo que respecta al BADYG Medio, las desigualdades descubiertas en la escala HMNV también permitieron observar que los escolares del grupo focal 1 (LF V LE CV) presentaban mejores puntuaciones en dicha escala que los del grupo de referencia (LF C LE C) ($t=-2.55$, $p\leq0.011$). En cuanto a la fiabilidad, ambas escalas presentaron buena consistencia interna en cada uno de los grupos objeto de estudio, oscilando sus coeficientes KR-20 entre 0.830 y 0.890.

En relación a la unidimensionalidad, en ambas escalas se pudo apreciar la predominancia de un primer componente, el cual explicaba una buena proporción de la varianza total del test, y un punto de inflexión que se situaba en torno al segundo componente, siendo el resto de los componentes mucho menores. No obstante, la obtención de estos resultados no fueron del todo concluyente a la hora de confirmar la unidimensionalidad del contructo evaluado por ambas escalas. Conviene destacar que la unidimensionalidad de la escala HMV para la muestra LF V LE V se realizó únicamente con 39 ítems, debido a que el ítem 11 de dicha escala fue eliminado por falta de variabilidad.

Detección del DIF

En términos generales, los resultados muestran un mayor predominio del funcionamiento diferencial de los ítems en la escala HMV, que en los ítems de la escala HMNV. Para la escala HMV, el estadístico MH detectó un total de 7 ítems con problemas de DIF. Ninguno para la comparación 1 [LF C LE C/LF V LE CV (BADYG Elemental)], 4 ítems para la comparación 2 [LF C LE C/LF V LE

CV (BADYG Medio)], de los cuales 2 ítems (I10, I19) presentaban DIF uniforme y los otros dos (I5, I38) DIF no uniforme, y 3 ítems con DIF uniforme (I10, I22 y I26) para la comparación 3 [LF C LE C/LF V LE V (BADYG Elemental)]. En cambio, la RL fue incapaz de detectar DIF en los 40 ítems de la citada escala.

Por lo que respecta a la escala HMNV, ésta no presentó problemas de funcionamiento diferencial en la primera y segunda comparación; pero por el contrario en la tercera comparación [LF C LE C/LF V LE V (BADYG Elemental)] el estadístico MH resultó significativo en uno de sus ítems, el ítem 38 con DIF no uniforme, y el estadístico de Wald en dos de ellos (I31 y I38), mostrando DIF uniforme y no uniforme, respectivamente.

En resumen, la obtención de estos resultados muestran con bastante claridad la ausencia de concordancia conseguida por los diferentes métodos de DIF utilizados. La única coincidencia a la hora de detectar la presencia de funcionamiento diferencial en los ítems tiene lugar al establecer la comparación 3 para la escala HMNV. Curiosamente, el ítem 38 de la escala HMNV es el único detectado con DIF no uniforme por ambos métodos estadísticos.

Análisis adicionales

Ante este hallazgo tan poco alentador, el siguiente paso fue averiguar qué aspectos podían contribuir a clarificar el comportamiento de tales métodos de análisis en la detección del DIF. Los análisis adicionales que siguieron fueron: a) examinar los parámetros *a* y *b* de los ítems detectados con DIF, y b) determinar la magnitud del DIF identificado. La estimación de los parámetros *a* y *b* de los ítems con DIF se obtuvo tras ajustar un modelo logístico de 3P. Respecto a la magnitud del DIF identificado, ésta se computó

Tabla 1
Resumen de los índices de DIF detectados en la escala HMV

	MH				*RL*			
	DIF U		*DIF NU*		*DIF U*		*DIF NU*	
	χ^2_{MH}	*p*	χ^2_{MH}	*p*	*Wald*	*p*	*Wald*	*p*
Comparación 1								
LF C LE C/LF V LE CV								
BADYG Elemental								
	—	—	—	—	—	—	—	—
Comparación 2								
LF C LE C/LF V LE CV								
BADYG Medio								
5	—	—	8.82	0.003	—	—	—	—
10	11.27	0.001	—	—	—	—	—	—
19	10.83	0.001	—	—	—	—	—	—
38	—	—	9.17	0.002	—	—	—	—
Comparación 3								
LF C LE C/LF V LE V								
BADYG Elemental								
10	8.04	0.003	—	—	—	—	—	—
22	10.15	0.001	—	—	—	—	—	—
26	7.99	0.003	—	—	—	—	—	—

Tabla 2
Resumen de los índices de DIF detectados en la escala HMNV

	MH				*RL*			
	DIF U		*DIF NU*		*DIF U*		*DIF NU*	
	χ^2_{MH}	*p*	χ^2_{MH}	*p*	*Wald*	*p*	*Wald*	*p*
Comparación 1								
LF C LE C/LF V LE CV								
BADYG Elemental								
	—	—	—	—	—	—	—	—
Comparación 2								
LF C LE C/LF V LE CV								
BADYG Medio								
	—	—	—	—	—	—	—	—
Comparación 3								
LF C LE C/LF V LE V								
BADYG Elemental								
31	—	—	—	—	2.79	0.005	—	—
38	—	—	11.49	0.000	—	—	-3.55	0.000

aplicando la medida de área exacta sin signo, propuesta por Raju (1988, 1990). El valor de los parámetros *a* y *b* fue el obtenido en el análisis anterior, mientras que el parámetro *c* fue idéntico para ambos grupos, siendo su valor de 0.12. En la siguiente tabla (ver tabla 5), podemos apreciar los resultados obtenidos tras realizar los análisis mencionados.

Para el parámetro *a*, los valores obtenidos se encuentran dentro del intervalo [0.363 y 0.923], siendo los más frecuentes aquéllos que oscilan entre 0.40 y 0.70. En cuanto al parámetro *b*, éstos se encuentran entre los límites [-1.593 y 2.915]. Teniendo en cuenta los parámetros conseguidos, conviene destacar que la RL detecta con funcionamiento diferencial aquellos ítems con valores elevados en dificultad (I31 y I38), mientras que el patrón de detección de DIF ofrecido por el estadístico MH parece ser independiente de los parámetros conseguidos en *a* y *b*. Respecto a la magnitud del DIF detectado, los resultados conseguidos no ofrecen ninguna pauta de actuación definida, ya que ambos métodos de análisis son capaces de detectar varios tamaños de DIF, oscilando sus magnitudes de 0.246 a 1.499.

Discusión

El objetivo principal de este estudio ha sido averiguar el grado de congruencia mostrado por el estadístico Mantel-Haenszel y la regresión logística en la detección del funcionamiento diferencial de los ítems en dos pruebas de aptitud intelectual, elaboradas y baremadas en castellano y de uso habitual en el ámbito educativo valenciano.

Los resultados obtenidos por los dos métodos de análisis indican una falta de concordancia considerable en la detección del DIF de las dos pruebas psicológicas analizadas (HMV y HMNV). Mientras el estadístico MH logró identificar un total de 8 ítems con DIF en las 3 comparaciones realizadas, la RL únicamente detectó 2 de ellos, y la correspondencia entre ambos métodos tan sólo se vió confirmada para el ítem 38 de la escala HMNV. Estos resultados fueron interpretados en los ítems con DIF atendiendo a los valores obtenidos en los parámetros *a* y *b*, y a la magnitud del DIF localizado. Tales resultados permitieron descubrir que los dos ítems detectados con DIF por la RL presentaban niveles altos en dificultad (I31 y I38); en cambio, el resto de las condiciones apenas contribuyó a esclarecer las razones del comportamiento mostrado por ambos métodos de análisis.

A fin de clarificar estos resultados, se revisaron los estudios de simulación realizados sobre los dos métodos de detección utilizados (Ferreres, Fidalgo y Muñiz, 1999; Fidalgo, 1996; Narayanan y Swaminathan, 1996; Rogers y Swaminathan, 1993). En líneas generales, los resultados de estos estudios muestran que cuando el ítem presenta DIF uniforme la potencia de prueba del estadístico MH disminuye a medida que aumenta la dificultad del ítem; por el contrario, cuando el DIF es no uniforme, los ítems peor identificados por el MH son los ítems de dificultad medio-baja y baja discriminación. Por su parte, la RL ofrece un comportamiento aceptable en la detección del DIF de los ítems con baja (b=-1.5) o moderada (b=0) dificultad y elevada discriminación (a=0.75). Sin embargo, los ítems peor identificados por ésta son los ítems con niveles medios de dificultad y baja discriminación. Así pues, en lo

Tabla 3
Parámetros *a* y *b*, y magnitud del DIF en los ítems de las escalas HMV y HMNV

HMV					
	GR. REFERENCIA		GR. FOCAL		
	b	*a*	*b*	*a*	ΔDIF
Comparación 2					
LF C LE C/LF V LE CV					
BADYG Medio					
I5	-1.593	0.367	-0.387	0.363	1.025
I10	-0.148	0.733	-0.911	0.628	0.648
I19	0.177	0.660	0.495	0.923	0.270
I38	1.386	0.591	2.348	0.555	0.817
Comparación 3					
LF C LE C/LF V LE V					
BADYG Elemental					
I10	-0.620	0.400	-0.910	0.676	0.246
I22	-0.369	0.462	0.693	0.468	0.902
I26	-0.265	0.698	0.139	0.639	0.343
HMNV					
	GR. REFERENCIA		GR. FOCAL		
	b	*a*	*b*	*a*	ΔDIF
Comparación 3					
LF C LE C/LF V LE V					
BADYG Elemental					
I31	2.274	0.866	1.703	0.877	0.485
I38	2.915	0.903	1.151	0.602	1.499

que respecta al estadístico MH, nuestros resultados se comportan, en la mayoría de los casos, como apuntan los estudios de simulación señalados. Sin embargo, esto no ocurre en el caso de la RL. Por lo tanto, los resultados obtenidos sugieren que el estadístico MH puede ser útil en la detección del DIF en condiciones similares a las del presente estudio.

En definitiva, este estudio evidencia la debilidad que supone la aplicación de un único método de detección de DIF en estudios aplicados. En este caso, la identificación de los ítems con DIF en las escalas HMV y HMNV no coincide entre los diferentes métodos utilizados. Así pues, la toma de decisiones acerca de la eliminación de aquellos ítems con DIF en estudios empíricos debe basarse en múltiples y mayores evidencias. Ante estos resultados, queda manifiesta la necesidad de realizar nuevos estudios de DIF con el fin de averiguar las causas o factores que provocan el funcionamiento diferencial de los ítems en los distintos grupos analizados. Precisamente, la identificación de estos factores permitirá a los constructores de tests un mayor conocimiento de las diferencias en procesos de percepción, cognición y contenido, que subyacen a la hora de responder los distintos ítems de un determinado test y, por consiguiente, prevenir con mayor facilidad la presencia del problema del DIF.

Por otra parte, conviene destacar que este trabajo de investigación presenta algunas limitaciones. La primera de ellas reside en el hecho de que es un estudio empírico y, por tanto, no se puede conocer la cantidad de falsos positivos (FP) y falsos negativos (FN) incluidos en los resultados obtenidos. La segunda se refiere al tamaño muestral utilizado. Una de las desventajas de los estudios empíricos es que se dispone de pocos sujetos en el grupo focal y, en ocasiones, el grupo de referencia tampoco resulta ser suficientemente numeroso. En nuestro estudio, el tamaño muestral utilizado ha sido el requerido para aplicar adecuadamente los métodos estadísticos MH y RL. No obstante, en próximos estudios sería conveniente ampliar el tamaño muestral de los distintos grupos objeto de estudio, y así poder utilizar otros métodos de detección de DIF, en especial los basados en la TRI; de esta forma, sería posible examinar el grado de congruencia existente entre los diferentes métodos de la TRI y los basados en tablas de contingencia.

Referencias

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, 149-166.

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

Candell, G. y Hulin, Ch.(1987). Cross-Language and Cross-Cultural comparisons in scale translations: Independent sources of information about item functioning. *Journal of Cross-Cultural Psychology,17*,4,417-440

Ellis, B.B. (1989). Differential Item Functioning: Implications for Test Translations. *Journal of Applied Psychology, 74*, 6, 912-921.

Ferreres, D. (1998). *Funcionamiento diferencial de los ítems de una prueba de aptitud intelectual en función de la lengua familiar y la lengua de escolarización.*Tesis doctoral no publicada. Universitat de València.

Ferreres, D.; González-Romá, V.; LLoret, S.; Gómez, J. *Tests y Bilingüismo.* La Investigación en Metodología de las Ciencias del Comportamiento:Logros y Perspectivas.València, 1999.

Fidalgo, A.M. (1994). MHDIF: A computer program for detecting uniform and nonuniform differential item functioning with the Mantel-Haenszel procedure. *Applied Psychological Measurement,18*, 3, 300.

Fidalgo, A.M.(1996). *Funcionamiento diferencial de los items: Procedimiento Mantel-Haenszel y modelos loglineales.* Tesis doctoral no publicada. Universidad de Oviedo.

Gómez, J. e Hidalgo, M.D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: Una revisión metodológica. *Anuario de Psicología, 74*, 3, 1997.

Holland, P.W. y Thayer, D.T. (1986). *Differential item functioning and the Mantel-Haenszel procedure*. (Technical Rep. No. 86-69). Princeton, NJ: Educational Testing Service.

Hulin, Ch.L. y Mayer, L.J. (1986). Psychometric Equivalence of a Translation of the Job Descriptive Index into Hebrew. *Journal of Applied Psychology, 71*, 1, 83-94

Mazor, K.; Clauser, B. y Hambleton, R.K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel Procedure. *Educational and Psychological Measurement, 54*, 2, 284-291.

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

Rogers, H.J. y Swaminathan, H. (1993). A comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement, 17*, 2, 105-116.

Swaminathan, H. y Rogers, H.J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement, 27*, 4, 361-370.

Waller, N.G. (1998). EZDIF: Detection of Uniform and Nonuniform Differential Item Functioning with the Mantel-Haenszel and Logistic Regression Procedures. *Applied Psychological Measurement, 22*, 2, 391.

Yuste, C. (1988). *B.A.D.Y.G.-Elemental y Medio.* Madrid. Ciencias de la Educación preescolar y especial.