Psicothema 2007. Vol. 19, nº 2, pp. 329-336
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG
Copyright © 2007 Psicothema

Aplicación de cuatro procedimientos de detección del funcionamiento diferencial sobre ítems politómicos

Paula Elosua y Alicia López-Jáuregui
Universidad del País Vasco

En este trabajo se describen cuatro procedimientos de detección del funcionamiento diferencial del ítem para formatos de respuesta politómica: Mantel, Mantel-Haenszel Generalizado (GMH), Regresión Logística Ordinal (RLO) y Regresión Logística Discriminante (RLD). Además de los modelos teóricos se describen las medidas del tamaño del efecto utilizables con cada uno de ellos. Con un diseño de validación cruzada se analizan los ítems politómicos de dos cuadernillos de la prueba de comprensión lectora del programa PISA2000. Las muestras analizadas provienen de Estados Unidos y España. Adoptando como regla de decisión la significación de la prueba estadística y la medida del tamaño del efecto, el acuerdo entre los procedimientos evaluados es total para dos de los ítems analizados.

Application of four procedures for detecting differential item functioning in polytomous items. The authors describe and use four methods for detecting Differential Item Functioning in polytomous items: Mantel, Generalized Mantel-Haenszel (GMH), Ordinal Logistic Regression (RLO), and Discriminant Logistic Regression (RLD). For each procedure, the theoretical model and the measure of effect size are described. The data from the «Reading Comprehension Test» from the PISA2000 evaluation program were analyzed using a cross-validation design. Two booklets were independently evaluated in the American and Spanish samples. Adopting as decision rule the significance of the statistical test and the measurement of the effect size, agreement among the evaluated procedures was total for two of the analyzed items.

El sesgo de los ítems es una de las mayores amenazas contra la validez de los tests psicológicos y educativos. La posibilidad de que independientemente al uso propuesto existan factores que añadan varianza no deseada a la puntuación obtenida en un test (sexo, raza, curriculo, experiencia con el formato de respuesta…) hace necesaria la utilización, de forma sistemática, de procedimientos para la detección del funcionamiento diferencia del ítem (FDI). La literatura está repleta de estudios sobre métodos de detección del FDI en ítems de respuesta dicotómica (Berk, 1982; Holland y Wainer, 1993; Camilli y Shepard, 1994; Fidalgo, Mellenbergh, y Muñiz, 1998; Elosua, López, y Egaña, 2000). Sin embargo, es menor el número de trabajos destinados a profundizar en los métodos de detección aplicables a formatos de respuesta ordenada o escala Likert.

Los procedimientos de detección de FDI en ítems dicotómicos comparan las respuestas dadas a un ítem por sujetos que provienen de dos grupos (referencia/focal) y tienen el mismo nivel en la variable medida (puntuación total o nivel de habilidad estimado). Los procedimientos aplicables a ítems politómicos son más complejos que los utilizados con formatos dicotómicos. Por un lado, el formato de respuesta ordinal tiene más categorías que el dicotómico, lo cual dificulta la comparación de las «respuestas dadas al ítem»; esta comparación podría llevarse a cabo teniendo en cuenta la media aritmética del ítem, o teniendo en cuenta las diferencias asociadas con cada una de las opciones de respuesta o sobre todas ellas conjuntamente (French y Miller, 1996). Por otro lado, la utilización de ítems de respuesta ordenada amplía el rango de puntuaciones utilizado para crear los niveles de habilidad necesarios para emparejar sujetos antes de poder ser comparados.

El objetivo de este trabajo es describir y aplicar cuatro procedimientos de detección del FDI sobre ítems con formato de respuesta ordinal: Mantel, Mantel-Haenszel Generalizado, Regresión Logística Ordinal y Regresión Logística Discriminante (Hidalgo y Gómez-Benito, 1999; Miller y Spray, 1993; Spray y Miller, 1993; Tian, 1999). Los dos primeros son procedimientos no-paramétricos basados en el análisis de tablas de contingencia. La regresión logística ordinal y la regresión logística discriminante son métodos paramétricos que evalúan la presencia de FDI por medio de la comparacion de modelos anidados.

Métodos de detección

Métodos Mantel

La familia de los estadísticos Mantel (1963; Mantel y Haenszel, 1959; Holland y Thayer, 1988) evalúan la asociación entre variables categóricas. En su aplicación al estudio del FDI analizan la asociación entre las respuestas dadas a un ítem y la pertenencia a

Fecha recepción: 3-10-06 • Fecha aceptación: 31-10-06
Correspondencia: Paula Elosua
Facultad de Psicología
Universidad del País Vasco
20018 San Sebastián (Spain)
E-mail: Paula.elosua@ehu.es

un grupo (grupo de referencia/grupo focal). Para ello se divide la puntuación total en varios intervalos diferentes (niveles de habilidad o niveles de puntuación; K) y se lleva a cabo la comparación entre los grupos en cada uno de ellos para obtener finalmente un estadístico sobre todos los niveles evaluados.

Para estudiar la asociación los datos se organizan en tablas de dimensiones $2\times M\times K$, donde M es el número de categorías de respuesta (M= 2 en el caso de respuestas dicotómicas 0/1) y K es el número de niveles en los que se ha dividido la puntuación total. En cada uno de los K niveles los datos se representan del siguiente modo (véase tabla 1).

Los valores $y_1, y_2, \ldots y_M$ representan las opciones de respuesta. El cuerpo de la tabla se completa con el número de personas provenientes del grupo de referencia (n_R) o grupo focal (n_F) que dentro del grupo de puntuación K obtienen la puntuación y_m. Los marginales de las tablas (+) representan las sumas de las filas o las columnas correspondientes.

Mantel

Es un estadístico propuesto por Mantel (1963) que evalúa la asociación entre filas/columnas (grupos/respuestas) a través de la comparación de las medias obtenidas en dos grupos una vez igualados en función de la variable de emparejamiento (puntuación total). El estadístico que estima la interacción entre los grupos y las categorías de respuesta se distribuye con 1 grado de libertad.

$$Mantel\ \chi^2 = \frac{\left(\sum_k F_k - \sum_k E(F_k)\right)^2}{\sum_k Var(F_k)}$$

Donde F_k es la suma de las puntuaciones obtenidas por el grupo focal en el nivel k de la variable de emparejamiento:

$$F_k = \sum_m y_m n_{Fmk}$$

El valor esperado de F_k y su varianza bajo la hipótesis nula de no asociación serían:

$$E(F_k) = \frac{n_{F+k}}{n_{++k}} \sum_m y_m n_{+mk}$$

$$Var(F_k) = \frac{n_{R+k} n_{F+k}}{n^2_{++k}(n_{++k}-1)} \left\{ \left(n_{++k} \sum_m y_m^2 n_{+mk} \right) - \left(\sum_m y_m n_{+mk} \right)^2 \right\}$$

Tabla 1
Representación de los ítems politómicos

	Categoría de respuesta					
Grupo	y_1	y_2	y_3	…	y_M	**Total**
Referencia	n_{R1k}	n_{R2k}	n_{R3k}	…	n_{RMk}	n_{R+k}
Focal	n_{F1k}	n_{F2k}	n_{F3k}	…	n_{RMk}	n_{F+k}
Total	n_{+1k}	n_{+2k}	n_{+3k}		n_{+Mk}	N_{++k}

Mantel-Haenszel Generalizado

Es una generalización del estadístico Mantel-Haenszel (Mantel y Haenszel, 1959; Spray y Miller, 1994; Tian, 1999; Zwick, Donogue, y Grima, 1993) para datos de respuesta nominal que analiza las diferencias entre los grupos a través de la comparacion de las distribuciones de las respuestas. Siguendo la notación de la tabla anterior, la estimación del estadístico vendría dada por:

$$\mathbf{A}'_k = \left(n_{R1k}, n_{R2k}, \ldots\ n_{R(M-1)k} \right)$$

$$E\left(\mathbf{A}'_k\right) = n_{R+k} \mathbf{n}_{k} / n_{++k}$$

$$\mathbf{n}'_k = \left(n_{+1k}, n_{+2k}, \ldots\ n_{+(M-1)k} \right)$$

$$\mathbf{V}(\mathbf{A}_k) = n_{R+k} n_{F+k} \left(\frac{n_{++k} diag(\mathbf{n}_k) - \mathbf{n}_k \mathbf{n}'_k}{n^2_{++k}(n_{++k}-1)} \right)$$

El test que evalúa la asociación es:

$$GMH\ \chi^2 = \left[\sum \mathbf{A}_k - \sum E(\mathbf{A}_k)\right]' \left[\sum \mathbf{V}(\mathbf{A}_k)\right]^{-1} \left[\sum \mathbf{A}_k - \sum E(\mathbf{A}_k)\right]$$

Este estadístico se distribuye con M-1 grados de libertad bajo la hipótesis nula de no asociación. Como puede comprobarse, este estadístico no considera el posible orden existente entre las categorías y compara las distribuciones de los grupos en un ítem sin tener únicamente en cuenta los valores medios.

Medida del tamaño del efecto

El tamaño del efecto puede analizarse a través de la diferencia entre medias estandarizadas (SMD; Dorans y Kulick, 1986; Zwick y Thayer, 1996). Este índice es una extensión de la formulación de Dorans y Holland (1993) que proponen como indicador de FDI la diferencia entre las medias de los grupos de referencia y focal. El nuevo estadístico cuantifica la diferencia entre la media obtenida en el grupo focal (minuendo) y la media del grupo de referencia «estandarizada» como si la distribución del grupo de referencia fuera la misma que la del grupo focal (sustraendo).

$$SMD = \sum_k \frac{n_{F+k}}{n_{F++}} \frac{\sum_m y_m n_{Fmk}}{n_{F+k}} - \sum_k \frac{n_{F+k}}{n_{F++}} \frac{\sum_m y_m n_{Rmk}}{n_{R+k}}$$

Un valor negativo indicaría que el ítem favorece al grupo de referencia.

Dado que el valor de este índice depende de la escala de respuesta, es posible normalizarlo dividiendo el valor obtenido (SMD) por la desviación estándar obtenida en el ítem combinando los grupos de referencia y focal. El nuevo estadístico es SMD/S_i. Siguiendo el criterio utilizado por la *Educational Testing Service* (ETS) en la clasificación del grado de severidad del FDI en ítems politómicos, para que un ítem presente FDI moderado además de la significación del estadístico utilizado (α= 0,05), el tamaño del efecto será mayor o igual que 0,17 y menor o igual que 0,25. El ítem presentará FDI severo si además de la significación estadística el tamaño del efecto es mayor que 0,25.

Regresión Logística Ordinal

El modelo de regresión logística para datos dicotómicos modela la probabilidad de respuesta correcta en función de la puntuación total (Total), la pertenencia al grupo (Grupo) y la interacción entre ambas variables (Grupo×Total).

$$\text{logit} = \ln\left(\frac{p(X_i=1)}{p(X_i=0)}\right) = b_0 + b_1 Total + b_2 Grupo + b_3 Grupo * Total$$

Para evaluar el FDI se comparan las razones de verosimilitud de los modelos anidados (total, total+grupo, total+grupo+interacción). El modelo base se construye únicamente respecto al parámetro de la variable que indica el nivel de habilidad (Total). La existencia de FDI uniforme se concluiría cuando la diferencia entre el modelo base y el modelo que incluye el parámetro de pertenencia al grupo (Total+Grupo) es significativa. El FDI no uniforme compara este segundo modelo con el modelo que incluye el término de interacción (Total+Grupo+Interacción). Para las situaciones de respuetas politómicas el modelo se extiende dando lugar a tres variaciones básicas que dependen de la definición de los logit: el modelo acumulativo, el modelo continuo y el modelo de categorías adyacentes (Agresti, 1984, 1990), siendo de entre todos ellos el más utilizado el modelo acumulativo. En este modelo se compara la probabilidad de que la respuesta al ítem (Y) sea menor o igual que la opción de respuesta *j*, con la probabilidad de que la respuesta (Y) sea mayor a la opción de respuesta j:

$$\log\left[\frac{P(y_m \le j)}{P(y_m > j)}\right] = \alpha_j + b_1 Total + b_2 Grupo + b_3 Total * Grupo$$

$$\text{logit}\left[P(y_m \le j)\right] = \alpha_j + b_1 Total + b_2 Grupo + b_3 Total * Grupo$$

Donde *j* indica la categoría de respuesta j= 1,2,…*m*

El procedimiento de detección de FDI es similar al caso dicotómico (Swaminathan y Rogers, 1990). Evalúa la presencia de FDI a través del estudio de la mejora en el ajuste que produce la incorporación sucesiva de los parámetros mencionados al modelo de regresión logística (Puntuación Total, Puntuación Total+Grupo, Puntuación Total+Grupo+Puntuación Total×Grupo). Este método, además de un test de significación basado en la diferencia entre las razones de verosimilitud de dos modelos anidados, incluye una medida del efecto del FDI. Esta medida está asociada a las diferencias en el estadístico R^2 de Nagelkerke entre dos modelos (Thomas y Zumbo, 1998). La medida, R^2, representa la proporción de variación de las respuestas al ítem explicada por el modelo de regresión. Un ítem presenta FDI moderado cuando la diferencia entre modelos es significativa y además el incremento en R^2 entre los modelos base y el que incorpora el término de interacción se sitúa entre los valores 0,035 y 0,070. Un ítem presenta un FDI notable cuando además de la significación del estadístico, la diferencia entre los R^2 es superior o igual a 0,070 (Jodoin y Gierl, 2001). Este valor incremental de R_2 puede descomponerse y analizarse para cada par de modelos anidados ($R^2_{Modelo2}$-$R^2_{Modelo1}$, y $R^2_{Modelo3}$-$R^2_{Modelo2}$) y obtener información sobre el tipo de FDI (Gelin y Zumbo, 2003).

Regresión Logística Discriminante (RLD)

Al igual que la regresión logística, este procedimiento basa la detección del FDI en la comparación de modelos. La mayor diferencia entre estos dos acercamientos es que la RLD modela la pertenencia al grupo en lugar de la probabilidad de respuesta (Miller y Spray, 1993; Spray y Miller, 1994; Tian, 1999). Es decir, la pertenencia al grupo se pronostica a partir de la puntuación total, la respuesta al ítem y la interacción entre ambos factores. Estas dos variables (puntuación total, respuesta al ítem) permiten definir tres ecuaciones anidadas (Puntuación Total, Puntuación Total+Ítem, Puntuación Total+Ítem+Puntuación Total×Ítem) que posibilitan evaluar tanto el FDI uniforme como el FDI no uniforme. Para ello se computa el estadístico de razón de verosimilitud para cada uno de los modelos (G^2), y se compara entre dos modelos para concluir presencia/ausencia FDI. La literatura no ha descrito todavía medidas del tamaño del efecto asociadas a este modelo. Formalmente, la RLD podría representarse del siguiente modo:

$$P(Grupo|Y_i, Total) = \frac{\exp(1-Grupo)(-b_0 - b_1 Y_i - b_2 Total - b_3 Y_i Total)}{\left[1+\exp(-b_0 - b_1 Y_i - b_2 Total - b_3 Y_i Total)\right]}$$

Método

Participantes

La muestra está formada por 2.205 estudiantes, todos ellos de 15 años de edad. La muestra de referencia proviene de los Estados Unidos (N_R= 843), y la muestra focal es la muestra española (N_F= 1362).

Instrumento

El test de comprensión lectora de la evaluación internacional PISA2000 es un banco compuesto por 141 ítems distribuidos en 9 cuadernillos. En este trabajo se analizan los ítems politómicos liberados pertenecientes a los cuadernillos 8 y 9. Estos cuadernillos están compuestos por los mismos 31 ítems, de los que analizamos los 5 ítems de respuesta ordenada (0-1-2). Los análisis se llevan a cabo de modo independiente en cada uno de los cuadernillos, de acuerdo a un diseño de validación cruzada que incremente la validez externa de la investigación. La distribución de estudiantes por cuadernillos y países puede consultarse en la tabla 2.

Tabla 2
Estadísticos descriptivos para las muestras de referencia y focal

	Cuadernillo	N	M.A.	D.T.	% varrianza primer componente	α
EE.UU.	8	435	18,54	9,24	29,3	0,896
	9	408	18,10	8,94	29,1	0,904
España	8	679	17,41	7,81	22,5	0,861
	9	683	19,38	7,80	23,6	0,871

M.A.= Media Aritmética; D.S.= Desviación Típica; α= Alpha de Cronbach

Resultados

Análisis preliminares

Los estadísticos descriptivos para cada uno de los grupos se presentan en la tabla 2. La diferencia de medias intra-países y entre-cuadernillos no es significativa en la muestra de referencia (t= 0,705; p= 0,480; ω^2= 0,0005) y sí lo es en la muestra española (t= -4,65; p<0,001), donde el tamaño del efecto estimado por el índice ω^2 es 0,014. La media aritmética obtenida en el cuadernillo 9 (M.A.= 19,38) es significativamente mayor que la obtenida en el cuadernillo 8 (M.A.= 17,41).

Las diferencias encontradas entre países para el mismo cuadernillo son significativas. Las diferencias asociadas al cuadernillo 8 favorecen al grupo de referencia (t= -2,11; p= 0,035), mientras que las diferencias encontradas en el cuadernillo 9 indican un mejor rendimiento de la muestra española (t= 2,40; p= 0,016) (véase tabla 2). La medida del tamaño del efecto en el grupo de referencia fue ω^2= 0,003, y en el grupo focal ω^2= 0,004. En el diagrama de cajas siguiente se muestran las distribuciones asociadas a estas variables (véase figura 1).

Los estadísticos descriptivos de cada uno de los ítems analizados y estimados de modo independiente en cada grupo/cuadernillo se muestran en la tabla 3.

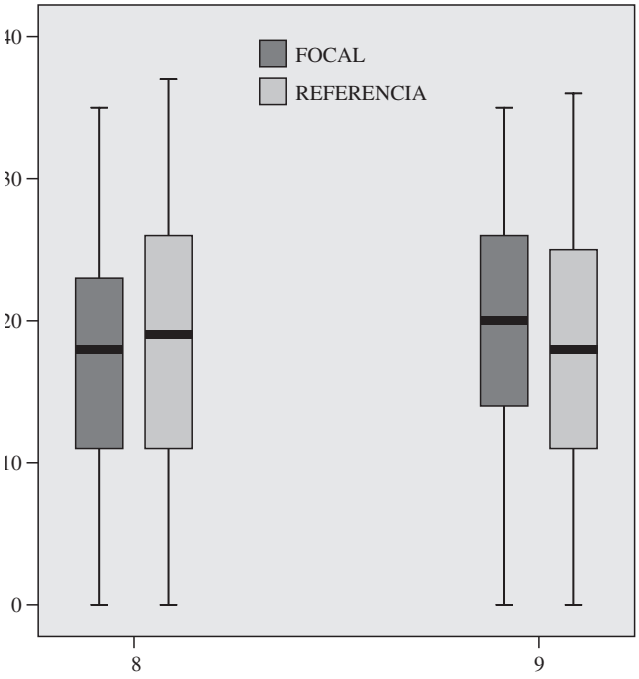

Figura 1. *Diagrama de cajas. Distribución de los grupos por cuadernillos*

Unidimensionaldad y consistencia interna

Se ha extraído un componente principal en cada una de las cuatro muestras analizadas. La tabla 2 recoge el porcentaje de varianza asociado a cada uno de estos componentes. En la muestra estadounidense el primer componente extraído explicó más del 29% de la varianza. En la muestra española este porcentaje es del 22,5% para el cuadernillo 8 y 23,6% para el cuadernillo 9. Los gráficos de sedimentación reflejan la estructura interna de cada uno de los cuadernillos (sólo se muestran los correspondientes al cuadernillo 8, porque los gráficos derivados del cuadernillo 9 son equivalentes).

El coeficiente alpha de Cronbach alcanza los valores de 0,896 y 0,904 en la muestra de referencia. Los valores obtenidos en la muestra española son 0,861 y 0,871 (tabla 2).

Análisis del funcionamiento diferencial del ítem

La aplicación de los procedimientos descritos se ha llevado a cabo en dos etapas. En la primera se estima el FDI utilizando la puntuación total sobre todos los ítems, y en la segunda se recalcula la puntuación total únicamente con los ítems libres de FDI; de este modo se intenta purificar la variable de emparejamiento (Holland y Thayer, 1988; Hidalgo y Gómez-Benito, 2003). Los análisis se han llevado a cabo en S-Plus. Los resultados se pueden consultar en las tablas 4 y 5.

Mantel

Atendiendo a los resultados del procedimiento Mantel, 3 ítems presentan valores significativos (α= 0,05) tanto en el cuadernillo 8 como en el cuadernillo 9 (R077Q03, R077Q05, R236Q02). Si la información sobre la significación es complementada con la información sobre el tamaño del efecto, concluiríamos que el FDI presente en el cuadernillo 8 para el segundo ítem es insignificante (R077Q05, SMD/Si= -0,14), para el ítem R077Q03 es severo (SMD/Si= 0,34) y para el ítem R236Q02 es moderado (SMD/Si= 0,17). En el cua-

Tabla 3
Estadísticos descriptivos para cada uno de los ítems analizados

	Grupo focal						Grupo referencia					
	Cuaderno 8			Cuaderno 9			Cuaderno 8			Cuaderno 9		
	M.A.	D.T.	r_{iX}	M.A.	D.T	r_{iX}	M.A.	D.T.	r_{iX}	M.A.	D.T.	r_{iX}
R077Q03	1,03	0,899	0,541	1,29	0,836	0,550	0,82	0,908	0,639	0,92	0,902	0,536
R077Q05	0,64	0,860	0,431	0.81	0,876	0,415	0,87	0,926	0,551	0,97	0,903	0,516
R088Q03	0,83	0,719	0,551	1,06	0,751	0,505	0,91	0,674	0,633	1,01	0,705	0,569
R088Q04	0,71	0,660	0,524	0,81	0,660	0,433	0,76	0,680	0,610	0,78	0,644	0,537
R236Q02	0,60	0,836	0,392	0,55	0,830	0,490	0,54	0,871	0,542	0,38	0,762	0,516

M.A.= Media Aritmética; D.T.= Desviación Típica

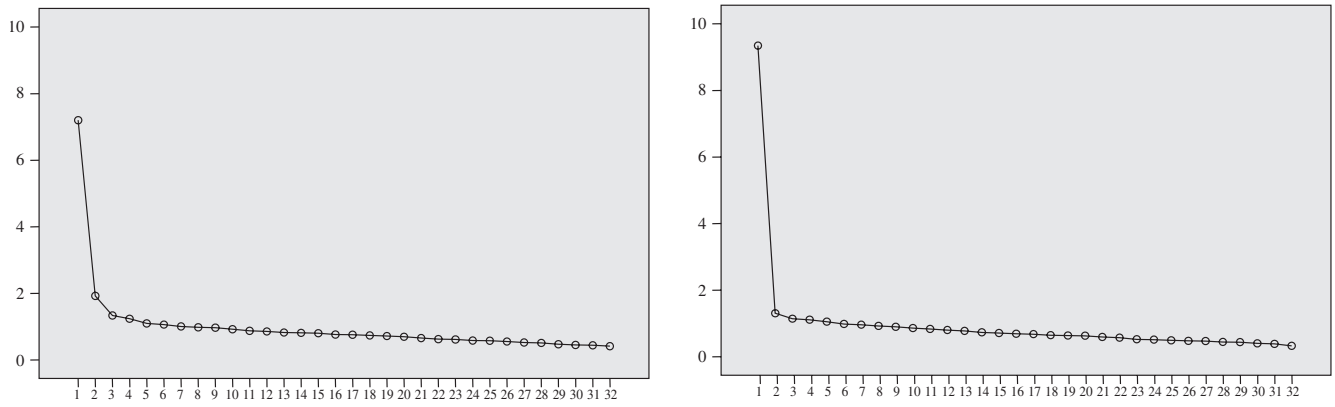

Figura 2. *Gráficos de sedimentación para las muestras focal y de referencia*

dernillo 9 el tipo de FDI encontrado es moderado para dos de los ítems (R077Q05, SMD/S_i= -0,19; R236Q02, SMD/S_i= 0,17) e intenso para uno de ellos (R077Q03, SMD/S_i= -0,31).

Los resultados conjuntos de los dos grupos de datos analizados muestran que tanto la clasificación de dos de los ítems (FDI/no FDI) y su grado de funcionamiento diferencial (insignificante/moderado/severo) es totalmente concordante para dos de los ítems; el ítem R077Q03 presenta un FDI severo y el ítem R236Q02 muestra un FDI moderado.

Mantel generalizado

Este procedimiento considera que los 5 ítems analizados en el cuadernillo 8 presentan funcionamiento diferencial. Este número se ve reducido en un ítem en los análisis llevados a cabo sobre el cuadernillo 9. Sin embargo, atendiendo a los valores mostrados por el estadístico utilizado para cuantificar el tamaño del efecto, concluiríamos que en el cuadernillo 8 la cantidad de FDI que presentan tres ítems es sustantivamente despreciable (R077Q05, RO88Q03, R088Q04). Los datos derivados del cuadernillo 9 nos llevarían a la misma conclusión para RO88Q03 y R088Q04, pues los valores SMD/S_i no alcanzan el mínimo fijado. Utilizando el criterio conjunto, significación estadística y tamaño del efecto, las conclusiones sobre la presencia de FDI serían las mismas que las relatadas para el procedimiento Mantel. En la interpretación de esta concordancia habría que tener en cuenta que ambos procedimientos utilizan la misma medida del tamaño del efecto.

Tabla 4
Resultados de los análisis de FDI del cuadernillo 8

Items	Mantel				SMD/S_i	Logística ordinal				Logística discriminante	
	Mantel		GMH			$G^2_{Mod.\ 2}$-$G^2_{Mod.\ 1}$	$G^2_{Mod.\ 3}$-$G^2_{Mod.\ 2}$	$R^2_{Mod.\ 2}$-$R^2_{Mod.\ 1}$	$R^2_{Mod.\ 3}$-$R^2_{Mod.\ 2}$	$G^2_{Mod.2}$-$G^2_{Mod.\ 1}$	$G^2_{Mod.\ 3}$-$G^2_{Mod.\ 2}$
	χ^2	p	χ^2	p							
R077Q03	**33,61**	**0**	**35,62**	**0**	**0,34**	**-32,45**	**-4,36**	**0,036**	**0,019**	**31,41**	**32,06**
R077Q05	6,83	0,00	6,88	0	-0,14	0,21	-0,21	0	0	**10**	**12,19**
R088Q03	0,22	0,63	12,18	0	-0,03	-1,22	-1,85	0	0,003	1,22	4,91
R088Q04	1,31	0,25	2,03	0	0,05	-0,29	-0,04	0	0	0,17	10,83
R236Q02	**5,91**	**0,01**	**24,69**	**0**	**0,17**	**6.72**	**11.06**	**0,02**	**0,073**	**1,38**	**32,07**

Tabla 5
Resultados de los análisis de FDI del cuadernillo 9

Items	Mantel				SMD/S_i	Logística ordinal				Logística discriminante	
	Mantel		GMH			$G^2_{Mod.\ 2}$-$G^2_{Mod.\ 1}$	$G^2_{Mod.\ 3}$-$G^2_{Mod.\ 2}$	$R^2_{Mod.\ 2}$-$R^2_{Mod.\ 1}$	$R^2_{Mod.\ 3}$-$R^2_{Mod.\ 2}$	$G^2_{Mod.2}$-$G^2_{Mod.\ 1}$	$G^2_{Mod.\ 3}$-$G^2_{Mod.\ 2}$
	χ^2	p	χ^2	p							
R077Q03	**36,07**	**0**	**38,14**	**0**	**0,31**	**-31,02**	**-0,73**	**0,035**	**-0,001**	**35,12**	**1,43**
R077Q05	**13,86**	**0**	**13,92**	**0,001**	**-0,19**	-17,13	-1,16	0,015	0,003	**20,01**	**7,46**
R088Q03	0,13	0,71	7,54	0,02	-0,02	-1,41	-2,48	0,002	0,003	0,86	3,22
R088Q04	0,021	0,88	1,72	0,42	0,014	-0,22	-0,16	0	0	0,23	6,2
R236Q02	**6,95**	**0,008**	**20,08**	**0**	**0,17**	**-9,28**	**-4,84**	**0,015**	**0,052**	**5,48**	**9,2**

Regresión Logística Ordinal

La diferencia de ajuste entre modelos es significativa para dos de los ítems en el cuadernillo 8 (R077Q03, R236Q02); la diferencia entre las R^2 de Nagelkerke para estos dos ítems supera el criterio de 0,035 fijado para determinar la presencia de FDI. En el primer caso la diferencia mayor se produce entre el ajuste del modelo base y del modelo que incorpora el parámetro de grupo ($G^2_{mod2\text{-}mod1}$= -32,45; $R^2_{mod2\text{-}mod1}$= 0,036), indicando la presencia de FDI uniforme. La diferencia mayor en el ajuste entre modelos para el ítem R236Q02 se produce entre el segundo modelo (total, grupo) y tercer modelo (total, grupo, interacción), indicando la presencia de FDI no-uniforme ($G^2_{mod3\text{-}mod2}$= -11,06; $R^2_{mod3\text{-}mod2}$= 0,073). En el cuadernillo 9 además de los dos ítems mencionados para el cuaderno 8 existe un tercer ítem que sería clasificado como FDI en función de las diferencias de ajuste entre modelos (R077Q05; $G^2_{mod2\text{-}mod1}$= -17,13); sin embargo, el tamaño del efecto para este ítem ($R^2_{mod2\text{-}mod1}$= 0,015) indicaría la presencia de un FDI insignificante.

Las conclusiones derivadas de la aplicación del modelo de regresión logística en los dos cuadernillos se resumirían en la concordancia total en la detección y clasificación de dos ítems, R077Q03 y R236Q02.

Regresión Logística Discriminante

Los resultados de la aplicación de este procedimiento en el cuadernillo 8 indicarían la presencia de FDI para tres ítems. El tipo de FDI de los dos primeros sería uniforme (R077Q03; $G^2_{mod2\text{-}mod1}$= 31,41; R077Q05; $G^2_{mod2\text{-}mod1}$= 10), mientras que el último presentaría FDI no uniforme (R236Q02; $G^2_{mod3\text{-}mod2}$= 32,07). Los resultados derivados del análisis del cuadernillo 9 son totalmente concordantes con los anteriores. Los mismos ítems presentan FDI en los dos cuadernillos.

De los resultados de los análisis efectuados por distintos procedimientos en dos conjuntos de datos independientes (cuadernillo 8 y cuadernillo 9) podría concluirse, utilizando un criterio conservador en la toma de decisión respecto al número de ítems que presentan FDI (Fidalgo, Ferreres, y Muñiz, 2004), que dos de los ítems analizados cumplen los criterios expuestos para la presencia de FDI por cada uno de los procedimientos utilizados. Estos dos ítems, R077Q03 y R236Q02, han sido detectados de modo independiente en los dos cuadernillos por el total de métodos empleados. El primero de ellos (R077Q03) presenta FDI uniforme y el segundo se ajusta a las características del FDI no-uniforme. La figura 3 muestra los valores medios para este ítem obtenidos en los grupos de referencia y focal estimados en 7 niveles de habilidad. El acuerdo sobre el grado de severidad del FDI encontrado no es total entre los procedimientos. El primer ítem presentaría un FDI severo de acuerdo a los procedimientos Mantel, sin embargo, sería moderado en función de las diferencias en las R^2 evaluadas con la RLO. El segundo ítem (R236Q02) sería clasificado como moderado por los métodos Mantel, sería severo de acuerdo a los análisis de regresión logística ordinal llevados a cabo en el cuadernillo 8 y moderado de acuerdo con los valores mostrados para este modelo en el cuadernillo 9.

Discusión y conclusiones

Son varios los trabajos que han evaluado la efectividad de los procedimientos de detección del FDI en formatos de respuesta ordenada. Todos ellos utilizan la simulación con un diseño en el que se manipulan factores como el tamaño del grupo, el parámetro de discriminación del ítem, el tipo de FDI, la cantidad de FDI o las diferencias entre las distribuciones de los grupos de referencia y focal (Tian, 1999; Spray y Miller, 1994; Zwick, Thayer, y Mazzeo, 1997: Kristjansson, Aylesworth, y Zumbo, 2005). En todos ellos se repiten las mismas conclusiones generales respecto a la potencia de los estadísticos en la detección y respecto al control de las falsas detecciones. A medida que aumenta el número de sujetos que componen las muestras se incrementa la efectividad de los procedimientos. En cambio, cuando el parámetro de discriminación del ítem estudiado es alto y además existen diferencias en la distribución de los grupos de referencia y focal la probabilidad de error tipo I supera el nivel nominal.

En el estudio que hemos mostrado en este artículo, el tamaño de los grupos es pequeño; no llega al número de 500 en la muestra de referencia (N_R= 435) y supera tímidamente esta cantidad en la muestra focal (N_F= 679); además existen diferencias significativas entre los grupos en los dos conjuntos de datos analizados (t= -2,11; p= 0,03; t= 2,40; p= 0,016), y los índices de discriminación de los ítems analizados son altos. Las condiciones empíricas que presentan nuestros datos no son las que se corresponden con las condiciones óptimas definidas para cada método; esta divergencia entre condiciones experimentales definidas en los estudios de simulación y condiciones empíricas derivadas del análisis de datos, por otro lado, es continua en la investigación psicométrica. En estas condiciones la utilización de diseños de validación cruzada como el empleado en este estudio es fundamental. Además de la aplicación conjunta de varios métodos de detección de FDI (Elosua, Ló-

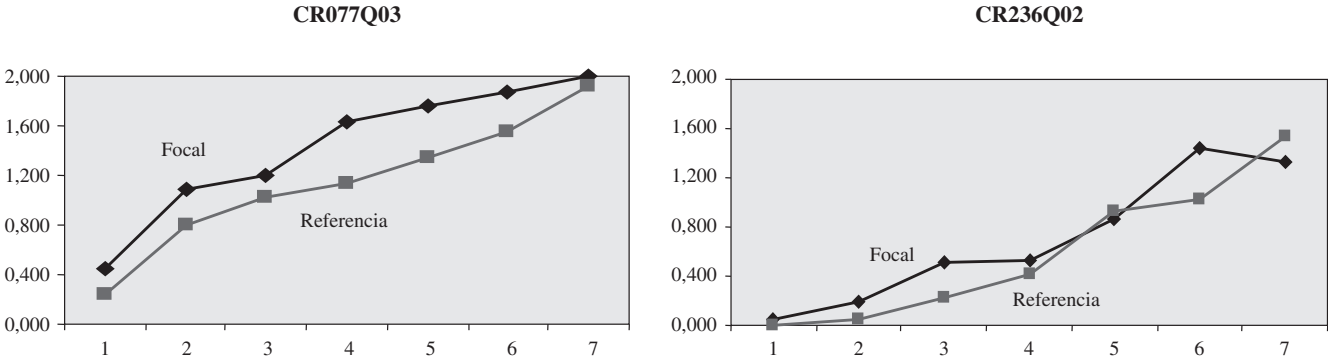

Figura 3. *Representación de las medias aritméticas del ítem obtenidas en 7 niveles de habilidad para los grupos de referencia y focal*

pez, y Torres, 2000; Fidalgo y Ferreres, 2002; Fidalgo, Ferreres, y Muñiz, 2004), siempre es aconsejable analizar conjuntos de datos independientes. La posibilidad de generalización de los resultados se amplía utilizando este diseño. Las conclusiones no serían idénticas analizando de modo independiente cada uno de los cuadernillos o utilizando un solo método de detección. Téngase en cuenta, por ejemplo, que los resultados de la prueba estadística asociada con el procedimiento Mantel-Haenszel generalizado han sido significativos para todos los ítems en uno de los cuadernillos.

Además de las consideraciones anteriores es importante resaltar la importancia de determinar el tamaño del efecto en los estudios de FDI. La necesidad de discriminar entre «resultado significativo» y «resultado sustantivamente relevante», y de neutralizar el papel determinante que el tamaño de la muestra tiene en el resultado de las pruebas de significación está impulsando de modo sistemático la utilización de medidas de tamaño del efecto en todas las áreas de investigación psicológica (Wilkinson y APA Task Force on Statistical Inference, 1999). En el campo del FDI se han desarrollado medidas para los procedimientos Mantel, tanto dicotómicos como politómicos, que han sido adoptadas por el *Educational Testing Service*. Las medidas de tamaño del efecto asociadas a la regresión logística ordinal están siendo objeto de investigación; son varias las propuestas existentes (Zumbo, 1999; Aylesworth y Kristjansson, 2000; Jodoin y Gierl, 200; Hidalgo y López, 2004), y ofrecen diferentes puntos de corte para los estadísticos propuestos. Sin embargo, existe todavía una indeterminación al respecto que exige una mayor labor de investigación en esta área. Los resultados empíricos mostrados en este trabajo concluyen que la concordancia entre las medidas del tamaño del efecto definidas para los procedimientos Mantel y para la regresión logística ordinal no ha sido absoluta. Téngase en cuenta que la determinación de los puntos de corte en la medida del tamaño del efecto es la base tanto para la detección de ítems anómalos como para la determinación de sus grados de severidad. En este trabajo hemos visto la influencia de ambas.

Agradecimientos

Trabajo financiado por el Ministerio de Educación (SEJ2005-01694).

Referencias

Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: Wiley and Sons.

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley and Sons.

Aylesworth, R., y Kristjanson, E. (2000). *Effect sizes measures for detecting polytomous DIF.* Unpublished working paper.

Berk, R.A. (ed.) (1982). *Handbook of methods for detecting item bias*. Baltimore: John Hopkins University Press.

Camilli, G., y L.A. Shepard (1994). *Methods for identifying biased test items.* London, Sage.

Dorans, N.J., y Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and Standarization. En P.W. Holland y H.Wainer (eds.): *Differential item functioning* (pp. 35-66) Hillsdale, NJ: Erlbaum

Dorans, N.J., y Kulick, E. (1986). Demonstrating the utility of the standarization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of educational measurement, 23*(4), 355-368.

Elosua, P., López, A., y Egaña, J. (2000). Idioma de aplicación y rendimiento en una prueba de comprensión verbal. *Psicothema, 12*(2), 201-206.

Elosua, P., López, A., y Torres, E. (2000). Desarrollos didácticos y funcionamiento diferencial de los ítems. Problemas inherentes a toda investigación empírica sobre sesgo. *Psicothema, 12*(2), 198-202.

Fidago, A.M., y Ferreres, D. (2002). Supuestos y consideraciones en los estudios empíricos sobre el funcionamiento diferencial de los ítems. *Psicothema*, *14*, 491-496.

Fidalgo, A.M., Ferreres, D., y Muñiz, J. (2004). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: Implications on the type I and type II error rate. *The Journal of Experimental Education, 73*(1), 23-39.

Fidalgo, A.M., Mellenbergh, G.J., y Muñiz, J. (1998). Comparación del procedimiento Mantel-Haenszel frente a los modelos loglineales en la detección del funcionamiento diferencial de los ítems. *Psicothema, 10* (1)*,* 219-228.

French, A.W., y Miller, T.R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315-332.

Gelin, M.A, y Zumbo, B.D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the centre for epidemiologic studies depression scale. *Educational and Psychological Measurement*, 63(1), 65-74.

Hidalgo, M.D., y Gómez-Benito, J. (1999). Técnicas de detección del funcionamiento diferencial en ítems politómicos. *Metodología de las Ciencias del Comportamiento*, *1*, 39-60.

Hidalgo, M.D., y Gómez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment, 19*, 1-11.

Hidalgo, M.D., y López, J.A. (2004). DIF detection and effect size: A comparision between logistic regression and Mantel-Haenszel Variation. *Educational and Psychological Measurement*, *64*, 903-915.

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

Holland, P.W., y Wainer, H. (eds.) (1993). *Differential item functioning*. Hillsdale, Lawrence Erlbaum Associates.

Jodoin, M.G., y Gierl, M.J. (2001). Evaluating type I error and power rates using an effect size with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14,* 329-349.

Kristjansson, E., Aylesworth, R., y Zumbo, B.D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935-953.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure*. Journal of American Statistical Association, 58*, 690-700.

Mantel, N., y Haenszel, W.M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Miller, T., y Spray, J. (1993). Logistic discriminant function analysis of DIF identification of polytomously scored items. *Journal of Educational Measurement*, *30*, 107-122.

Spray, J., y Miller, T. (1994). Identifying nonuniform DIF in polytomously scored test items (American College Testing Research Report Series 94-1). Iowa City, IA: American College Testing Program.

Swaminathan, H., y Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of educational measurement 27*(4), 361-370.

Tian, F. (1999). *Detecting differential item functioning in polytomous items*. Unpublished doctoral dissertation, Faculty of Education, University of Ottawa.

Thomas, D.R., y Zumbo, B.D. (1998). *Variable importance in logistic regression based on partitioning and R-squared measure.* Paper presented at the annual meeting of the Psychometric Society, Urbana, IL.

Wilkinson, L., y APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604

Zumbo, B.D. (1999*). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modelling as a unitary framework for binary and Likert-type item scores*. Ottawa, Canadá: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., J.R., Donogue, J., y Grima, K.L. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement 30*(3), 233-251.

Zwick, R., y D.T. Thayer (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics 21*(3), 187-201.

Zwick, R., Thayer, D.T., y Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*, 321-344.