Psicothema, 1998. Vol. 10, nº 1, pp. 209-218
ISSN 0214 - 9915 CODEN PSOTEG

COMPARACIÓN DEL PROCEDIMIENTO MANTEL-HAENSZEL FRENTE A LOS MODELOS LOGLINEALES EN LA DETECCIÓN DEL FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS

Ángel M. Fidalgo*, Gideon J. Mellenbergh** y José Muñiz*
* Universidad de Oviedo y ** Universidad de Amsterdam

Se ha realizado un estudio de simulación para comparar la eficacia del procedimiento Mantel-Haenszel y los modelos loglineales en la deteccion de ítems que funcionen diferencialmente. Dichos procedimientos se aplicaron en un primer y único análisis e iterativamente. Las principales conclusiones son: Primera, que el procedimiento Mantel-Haenszel tiene mayor potencia de prueba que los modelos loglineales. Segunda, que la aplicación iterativa de los procedimientos mejora sustancialmente las tasas de detección y el error tipo I en comparción con los correspondientes procedimientos no iterativos.

Comparison of the Mantel-Haenszel procedure versus the Loglinear models for detecting differential item functioning. A simulation study was carried out to compare the Mantel-Haenszel and the Loglinear procedures for detecting differentially functioning ítems. The procedures were applied in one single step and iteratively. The two main conclusions are: First, the Mantel-Haenszel procedure has higher power than the Loglinear procedures. Second, the iterative application of the procedures substantially improves detection rates and Type I error compared to single-step applications.

Dadas las implicaciones de carácter ético, social y jurídico que se derivan de la utilización de test o ítems sesgados contra algún grupo de la población, no es de extrañar la proliferación de investigaciones sobre el funcionamiento diferencial de los ítems (Fidalgo, 1996a). Se dice que un ítem funciona diferencialmente si examinados con el mismo nivel de competencia en la variable que pretende medir el test tienen diferentes probabilidades de responder correctamente a ese ítem. Además dentro del funcionamiento diferencial de los ítems cabría distinguir entre DIF uniforme y no uniforme (Mellenbergh, 1982). El DIF uniforme se produce cuando la probabilidad de contestar correctamente a un ítem es mayor para un grupo que para otro a través de todos los niveles de habilidad. El DIF no uniforme se produce cuando la diferencia en la probabilidad de responder correctamente a un ítem entre dos grupos no es la misma en todos los niveles de habilidad. Estas definiciones son unáni-

Correspondencia: Ángel Fidalgo
Universidad de Oviedo
Facultad de Psicología
Plaza de Feijoo, s/n. 33003 Oviedo (Spain)
E-mail: fidalgo@pinon.ccu.uniovi.es

memente aceptadas. Hay acuerdo en qué es el DIF, menos consenso existe en cómo detectarlo. Entre la variedad de métodos propuestos para detectar el DIF cabría destacar los métodos basados en el análisis de tablas de contingencia, tales como ji-cuadrado (Marascuilo y Slaugther, 1981), Mantel-Haenszel (Holland y Thayer, 1988), modelos loglineales, logit y de clase latente (Fidalgo, 1996b; Fidalgo y Mellenbergh, 1995; Fidalgo, Muñiz y Mellenbergh, 1997a; Fidalgo y Paz, 1995a; Fidalgo y Paz, 1995b; Kelderman, 1989; Kok, Mellenbergh y Van der Flier, 1985; Mellenbergh, 1982; Van der Flier, Mellenbergh, Adèr y Wijn, 1984), regresión logística (Swaminathan y Rogers, 1990) y los métodos basados en la teoría de respuesta a los ítems (TRI) (Hambleton y Rogers, 1989; Kim y Cohen, 1991; Lord, 1980; Muñiz, 1997; Raju, 1988, 1990). El principal inconveniente de los métodos TRI es que requieren grandes tamaños de muestra. Por contra, el procedimiento Mantel-Haenszel (MH) es particularmente atractivo por su fácil cálculo, la posibilidad de aplicarlo en muestras pequeñas, y el hecho de proporcionar una cuantificación del DIF presente en los ítems además de un test de significación estadística. Como se ha señalado, otra de las técnicas aplicadas en la detección del DIF que sí controlan el nivel de competencia de los sujetos son los modelos lineales logarítmicos o loglineales. Además , para conseguir que en lo posible el nivel de habilidad estimado a los sujetos no esté distorsionado por los ítems sesgados, se han establecido procedimientos de estimación en etapas. La lógica que subyace a todos ellos es evaluar el DIF en la forma estándar en una primera etapa, para en una segunda volver a evaluar el DIF, pero utilizando para estimar el nivel de competencia de los sujetos sólo aquellos ítems que no presentaron DIF en el análisis inicial (Lord, 1980). Distintos estudios de simulación han demostrado que la aplicación iterativa de los diversos procedimientos presenta siempre un mejor resultado que la aplicación de dichos procedimientos cuando el nivel de habilidad de los sujetos no se calculó sobre el test purgado (Fidalgo, 1996b; Fidalgo y Mellenbergh, 1995; Fidalgo, Muñiz y Mellenbergh, 1997b; Fidalgo y Paz, 1995a; Fidalgo y Paz, 1995b; Gómez y Navas, 1996; Hidalgo, 1995; Kok, Mellenbergh y Van der Flier, 1985; Miller y Oshima, 1992).

En este contexto, el presente estudio pretende, en primer lugar, determinar utilizando datos simulados la eficacia relativa del procedimiento MH frente a los modelos loglineales en la detección del DIF uniforme, y en segundo lugar, comprobar como afecta a ambos procedimientos su aplicación de forma iterativa.

Método

Variables

Las variables manipuladas en esta investigación han sido:

1. Cantidad de DIF, definida como la diferencia entre grupos en los índices de dificultad de los ítems que funcionan diferencialmente.
2. Tamaño de la muestra (75, 200, 500 y 1,000 sujetos por grupo).

Tenemos por tanto 16 condiciones, resultado de cruzar la variable Cantidad de DIF (factor intra) dentro de cada Tamaño de muestra (factor entre). Dentro de cada nivel del factor entre se llevaron a cabo 100 replicaciones para obtener resultados estables.

Generación de los datos

Se generaron vectores de respuestas (I_1, I_2,...,I_{40}), para diferentes tamaños de muestra, en un test compuesto por cuarenta ítems

dicotómicos a partir de un modelo unifactorial estricto. El modelo es el que sigue:

$$X_i = a_i\ F + u_i\ E_i \quad (a_i^2 + u_i^2 = 1),\ i = 1, 2, ..., 40 \qquad (1)$$

Siendo el término a_i la saturación factorial de ítem *i* en el factor común *F* , y el término u_i la saturación factorial del ítem *i* en el factor único E_i, considerado como término error en el modelo. Las puntuaciones factoriales (*F*) y las puntuaciones error (E_1, E_2, ..., E_{40}) son muestras independientes obtenidas de una distribución normal N(0, 1). Dado que tanto las puntuaciones factoriales como las error son independientes, las puntuaciones obtenidas por los sujetos en cada ítem (X_1, X_2, ..., X_{40}) también se distribuirán normalmente N(0, 1).

Las respuestas a los ítems I_1, I_2, ..., I_{40} se obtienen dicotomizando las puntuaciones obtenidas en las variables X_1, X_2, ..., X_{40} de la siguiente forma:

$$I_i = 0 \text{ si } X_i \leq z \text{ e } I_i = 1 \text{ si } X_i > z\,, \qquad (2)$$

siendo los puntos de corte elegidos (*z*, puntuaciones típicas) una función del índice de dificultad de los ítems (ID). Así

$$1 - ID = P(Z \leq z) \qquad (3)$$

De esta forma podemos manipular los índices de dificultad de los diferentes ítems que componen el test.

La longitud del test se fijó en 40 ítems, de los cuales el 10% presentaban sesgo. Los puntos de corte que determinan el índice de dificultad de los distintos ítems fueron elegidos, siguiendo a Lim y Drasgow (1990), dentro de un rango razonable entre -2 y +2, de tal forma que la distribución de las puntuaciones de los sujetos en el test se aproximen a una distribución normal. Los índices de dificultad de la totalidad de los ítems que componen el test se presentan en la Tabla 1.

Tabla 1
Valores paramétricos de los índices de dificultad de los ítems (ID), junto con las correspondientes puntuaciones *z*. Los ítems señalados presentan sesgo, siendo sus ID en el grupo 2: ítem 1= .40 (*z*= 0.25); ítem 2= .31 (*z*= 0.50); ítem 3= .16 (*z*= 1); ítem 4= .07 (*z*= 1.5)

Ítem	z	ID	Ítem	z	ID	Ítem	z	ID	Ítem	z	ID
1*	0.0	.50	11	-1.0	.84	21	0.0	.50	31	-1.0	.84
2*	0.0	.50	12	1.0	.16	22	0.0	.50	32	1.0	.16
3*	0.0	.50	13	1.5	.07	23	0.0	.50	33	1.5	.07
4*	0.0	.50	14	-1.5	.93	24	0.0	.50	34	-1.5	.93
5	0.5	.31	15	1.5	.07	25	0.5	.31	35	1.5	.07
6	-0.5	.69	16	-1.5	.93	26	-0.5	.69	36	-1.5	.93
7	0.5	.31	17	-2.0	.97	27	0.5	.31	37	-2.0	.97
8	-0.5	.69	18	2.0	.03	28	-0.5	.69	38	2.0	.03
9	-1.0	.84	19	-2.0	.97	29	-1.0	.84	39	-2.0	.97
10	1.0	.16	20	2.0	.03	30	1.0	.16	40	2.0	.03

Siguiendo el procedimiento especificado simulamos los datos para dos muestras independientes de sujetos, siendo la misma estructura factorial para todos los ítems en ambos grupos, y manipulando el índice de dificultad que diferirá, para los ítems con DIF, entre los grupos. La estructura factorial utilizada es

$$X_i = 0.80\ F + 0.60\ E_i \qquad (4)$$

La generación de los datos de acuerdo con el modelo y mediante el procedimiento expuesto se hizo dentro del paquete estadístico SPSS/PC+ V4.0, a partir de la posibilidad que brinda de generar distribuciones normales. De esta forma la cantidad de DIF, definida como la diferencia entre grupos en los índices de dificultad de los ítems que funcionan diferencialmente, fue igual a d_1 = .1, d_2 = .19, d_3 = .34, d_4 = .43, para el ítem primero, segundo, tercero y cuarto, respectivamente. Con objeto de expresar la magnitud del DIF que presentan estos ítems en términos del área existente entre la curva característica del ítem (CCI) para cada uno de los grupos, hay que tener en cuenta las relaciones existentes entre el parámetro *b* en

un modelo de TRI de un parámetro y un modelo unifactorial como el empleado en la generación de los datos (Ecuación 1). Esta relación viene dada por la siguiente expresión (Ferrando, 1994): $b_i = z_c / a_i$, siendo, el término a_i la saturación factorial de ítem *i* en el factor común *F* y z_c, el punto de corte elegido para dicotomizar. Así, para los cuatro ítems sesgados el valor del parámetro *b* en el grupo primero es igual a cero en todos los casos (0.0/0.8 = 0.0). Y el valor del parámetro *b* en el grupo segundo es igual a 0.3125, 0.625, 1.25 y 1.875, para los ítems primero, segundo, tercero y cuarto, respectivamente. Ahora que tenemos el valor de parámetro *b* en ambos grupos, y sabiendo que el DIF es uniforme, podemos aplicar la medida exacta del área con signo (Raju, 1988), obteniendo unas magnitudes de DIF contra el grupo segundo de 0.3125, 0.625, 1.25 y 1.875 para los ítems primero, segundo, tercero y cuarto, respectivamente.

Análisis

Los modelos loglineales permiten determinar las relaciones existentes entre una serie de variables categóricas representadas en tablas de contingencia multidimensionales. Este tipo de análisis especifican los parámetros que representan las propiedades de las variables categóricas y sus relaciones mediante la descomposición lineal de los logaritmos naturales de las frecuencias esperadas en una tabla de contingencia (Pardo y San Martín, 1994; Reynols, 1977). En el modelo saturado los componentes que definen la parcelación incluyen todos los efectos principales y las posibles interacciones entre las variables consideradas. En nuestro caso tenemos una tabla de contingencia multidimensional del tipo Puntuación x Grupo x Respuesta, con frecuencias observadas f_{ijk}, donde la variable Puntuación en el test tiene ocho categorías (i = 1, 2, ... , 8), la variable Grupo tiene dos categorías (j = 1 para el grupo de referencia y j = 2 para el grupo focal), y la variable Respuesta al ítem tiene dos categorías (k = 1 cuando se acierta el ítem y k = 2 cuando se falla).

Para determinar cuál es el modelo loglineal que mejor ajustaba, en una investigación anterior se utilizó la opción *Backward* dentro del comando *Hiloglineal* del programa estadístico SPSS/PC+ V4.0 (Fidalgo y Paz, 1995a). Para llevar a cabo estos análisis se toma la puntuación total del sujeto en el test como un índice del nivel de competencia del sujeto, y en función de ella los sujetos son asignados a cada uno de los diferentes niveles de habilidad establecidos (ocho intervalos de amplitud cinco). El procedimiento *Backward*, partiendo del modelo saturado, va eliminando parámetros y comprobando si el efecto de estos parámetros es estadísticamente significativo, es decir, si su presencia contribuye a mejorar el ajuste del modelo. Si un parámetro no es estadísticamente significativo es eliminado del modelo. Así se procede sucesivamente hasta conseguir un modelo en el que todos los términos de orden superior sean estadísticamente significativos (α = .05, en este estudio). En nuestro caso, concluiremos que existe DIF si en la clase generadora del modelo que mejor ajusta está presente la interacción entre la Respuesta al ítem y el Grupo. Es decir, si para el mismo nivel de habilidad existen diferencias entre los grupos en el número de sujetos que responden correctamente el ítem. Señalar que como índice de bondad de ajuste de los modelos se empleó la razón de verosimilitud ji-cuadrado (G^2).

Los modelos loglineales no distinguen entre variables dependientes e independientes. Es esta distinción, sin embargo, la que diferencia a un modelo logit de un modelo loglineal. Este tipo de modelos pueden ser considerados como un caso especial de modelos loglineales, de tal forma que cada modelo logit puede ser reformulado y tiene su equivalente modelo loglineal, e inversamen-

te. Por ejemplo, los modelos logit que indicarían la ausencia de DIF, el DIF uniforme y el DIF no uniforme son en ese orden:

$$\ln (F_{ij1} / F_{ij2}) = C + P_i , \quad (6)$$
$$\ln (F_{ij1} / F_{ij2}) = C + P_i + G_j , \quad (7)$$
$$\ln (F_{ij1} / F_{ij2}) = C + P_i + G_j + PG_{ij} , \quad (8)$$

donde, C es una constante, P_i denota el efecto del *í*-ésimo nivel de puntuación en el test, G_j el efecto del *j*-ésimo grupo, y PG_{ij} el efecto de la interacción del *í*-ésimo nivel de puntuación en el test y el *j*-ésimo grupo sobre el logaritmo neperiano de las razones esperadas de la variable dependiente, en este caso, la razón entre las respuestas correctas e incorrectas al ítem. Teniendo en cuenta esto, para determinar la importancia que la purificación de la puntuación en el test tiene en este tipo de procedimientos, se han comparado los resultados de aplicar el método iterativo logit (Van der Flier, Mellenbergh, Adèr y Wijn, 1984) con los resultados obtenidos en una investigación anterior (Fidalgo y Paz, 1995a), en la que se aplicaron los modelos loglineales en un único e inicial análisis a datos simulados bajo las mismas condiciones que las utilizadas en el presente estudio.

Descripción del método iterativo logit. El programa comienza dividiendo las puntuaciones de los examinados en el test en un número predeterminado de categorías. A continuación, en el primer paso, se determinan qué ítems del test ajustan al modelo logit correspondiente a la Ecuación 6. El ítem con un valor en el estadístico G^2 más alto y significativo es identificado. En el segundo paso el ítem identificado como sesgado en el primer análisis es excluido en el cálculo de las puntuaciones de los sujetos en el test. Se vuelven a formar las categorías de acuerdo con la puntuación purificada en el test (*n* - 1 ítems) y el modelo de la Ecuación 6 se vuelve a probar para todos los ítems. Los dos ítems con los valores en G^2 más altos y significativos son identificados. En un nuevo paso esos dos ítems son eliminados y las categorías son reordenadas en función de las puntuaciones en el test de *n* - 2 ítems, y el modelo se vuelve a ajustar para todos los *n* ítems. El procedimiento termina cuando excede el número de iteraciones fijado con anterioridad o cuando todos los ítems en el test tienen valores en G^2 no significativos. El número de categorías de puntuación utilizados fue, como en los modelos loglineales, de ocho. Además, se fijó un número elevado de iteraciones (40) de forma que el programa termine cuando todos los ítems en el test tienen valores en el estadístico G^2 no significativos

Descripción del programa MHDIF. El programa MHDIF (Fidalgo, 1994) fue utilizado para analizar cada base de datos. El programa implementa el procedimiento bietápico propuesto por Holland y Thayer (1988). El número de categorías que utiliza el programa para el cálculo del estadístico ji-cuadrado MH es igual al número de posibles puntuaciones en el test, es decir, 0, 1, ... , 40. Una vez calculado el estadístico ji-cuadrado MH, elimina los ítems con valores significativos en dicho estadístico, y recalcula de nuevo los estadísticos MH usando la puntuación total del sujeto en los ítems restantes como variable de bloqueo. Cuando un ítem está siendo investigado se incluye en la variable de agrupamiento aunque haya presentado DIF en el análisis inicial.

Resultados

Se ha tomado como medida de la eficacia del método iterativo logit y del procedimiento MH, tanto el porcentaje de ítems con DIF correctamente detectados, como el porcentaje de ítems sin DIF incorrectamente detectados (falsos positivos). Estos resultados, junto con los obtenidos al aplicar los modelos loglineales a datos simulados según el mismo modelo y bajo las mismas

condiciones (Fidalgo y Paz, 1995a), se pueden ver en la Tabla 2.

Tabla 2
Porcentajes de detecciones correctas y falsos positivos para los modelos loglineales (ML), el método iterativo logit (LOGIT), el MH aplicado en una única etapa (MH) y aplicado en dos etapas (MH-B). El nivel de significación utilizado en los análisis fue el .05

	Tamaño de muestra							
	75				200			
	ML	LOGIT	MH	MH-B	ML	LOGIT	MH	MH-B
Falsos p. ítems DIF	6.5	0.17	3.42	1.36	11	0.94	11.17	3.44
nº4	100	100	100	100	100	100	100	100
nº3	100	96	99	100	100	100	100	100
nº2	72	18	39	68	98	97	95	100
nº1	12	1	7	14	44	35	19	52
Total	71	53.75	61.25	70.5	85.5	83	78.5	88

	Tamaño de muestra							
	500				1.000			
	ML	LOGIT	MH	MH-B	ML	LOGIT	MH	MH-B
Falsos p. ítems DIF	25	1.50	28.13	6.27	45	1.19	45.03	9.03
nº4	100	100	100	100	100	100	100	100
nº3	100	100	100	100	100	100	100	100
nº2	100	100	100	100	100	100	100	100
nº1	67	77	42	91	95	100	81	98
Total	91.75	94.25	85.50	97.75	98.75	100	95.25	99.50

En la Figura 1 se representa gráficamente el porcentaje medio de detecciones correctas de los 4 ítems sesgados (los "totales" de la Tabla 2) que corresponde a cada método. Por su parte, el porcentaje de falsos positivos de cada método puede verse en la Figura 2.

Como se puede observar en la Figura 1, el procedimiento MH bietápico se muestra más eficaz en la detección de los ítems con DIF que el método iterativo logit, con independencia del tamaño de muestra. Por el contrario, el método iterativo logit presenta, en general, un menor número de falsos positivos que el procedimiento MH (Figura 2). Se hipotetizó que este último resultado puede deberse a que el método iterativo logit elimina sucesivamente a los ítems que presentan DIF en los análisis anteriores, obteniendo de esta forma una mejor estimación del nivel de habilidad de los examinados, y que, por tanto, si el procedimiento MH se aplicase no sólo una segunda vez sino iterativamente, el número de falsos positivos se reduciría aproximándose al del método iterativo logit. Para comprobar este extremo se realizaron 20 simulaciones para un tamaño de muestra de 1,000 sujetos por grupo, y se elaboró un programa que calculase el procedimiento MH de forma iterativa. El citado programa procede de la siguiente forma:

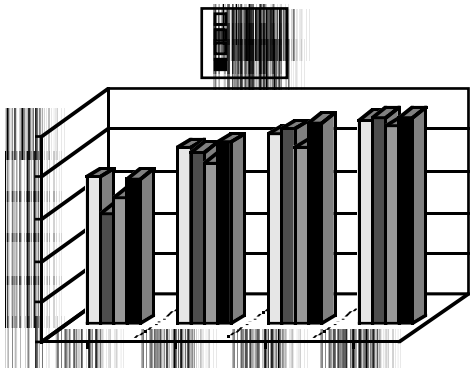

Figura 1. Porcentaje de identificaciones correctas para cada método y en cada tamaño de muestra al nivel de confianza del 95%.

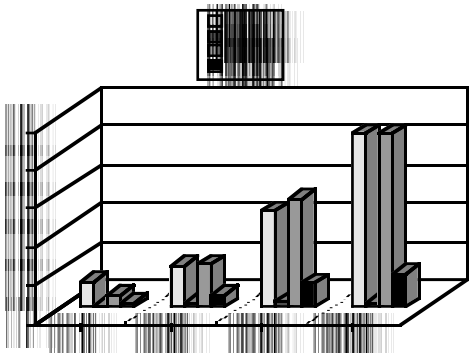

Figura 2.Porcentaje de falsos positivos para cada método en cada tamaño de muestra al nivel de confianza del 95%.

calcula estadístico ji-cuadrado MH utilizando un número de categorías de puntuación igual al número de posibles puntuaciones en el test. Elimina los ítems con valores significativos en el estadístico ji-cuadrado MH , y recalcula de nuevo los estadísticos MH usando la puntuación total del sujeto en los ítems restantes como variable de bloqueo. Así procede iterativamente hasta que (a) los ítems identificados con DIF son los mismos que los identificados en la iteración anterior, o (b) se llega al número máximo de iteraciones fijado (15 en este estudio). Siempre que un ítem está siendo investigado se incluye en el criterio de agrupamiento. Los resultados de dichos análisis se ofrecen en la Tabla 3. Los mismos confirman la hipótesis planteada, como puede observarse de forma evidente en la Figura 3.

Tabla 3
Porcentajes de ítems con DIF correctamente identificados y de falsos positivos por el procedimiento MH aplicado en una etapa (MH), aplicado en dos etapas (MH-B) e iterativamente (MH-I). El nivel de significación utilizado en los análisis fue el .05

	Método		
	MH	**MH-B**	**MH-I**
Falsos positivos	44.86	10.28	3.89
ítems con DIF			
nº4	100	100	100
nº3	100	100	100
nº2	100	100	100
nº1	95	100	100
Total	98.75	100	100

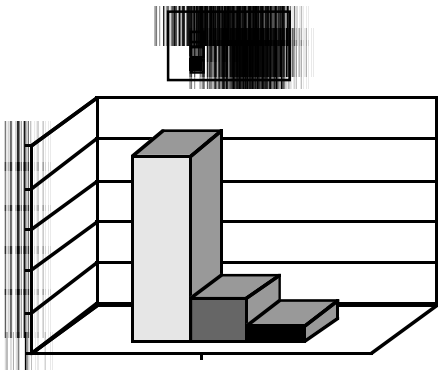

*Figura 3.*Porcentajes de falsos positivos para cada método de análisis con α = .05 y un tamaño de muestra de 1,000 examinados por grupo.

Discusión

La principal diferencia entre los modelos loglineales utilizados y el método iterativo logit es precisamente esa, que el segundo procedimiento opera iterativamente, utilizando en la construcción de las tablas de contingencia las puntuaciones purificadas de los examinados en el test. Entendemos que las diferencias entre los resultados de uno y otro se deben primordialmente a este hecho, aunque la opción *Backward* del SPSS parta del modelo saturado buscando otro más parsimonioso y con buen ajuste, en tanto que el procedimiento iterativo logit comprueba sólo el ajuste de los datos al modelo logit que especifica la no existencia de DIF.

El incremento en las tasas de detección conforme aumenta el tamaño de muestra era de esperar. Como es bien sabido cualquier estadístico aumenta su potencia de prueba, permaneciendo igual otras condiciones, conforme aumenta el tamaño de muestra. En trabajos anteriores (Fidalgo y Paz, 1995a) se hipotetizó que las elevadas tasas de error tipo I encontradas al aplicar los modelos loglineales (desde .065 en el caso más favorable hasta un .45 en el peor, con un α= .05) podían deberse a que la puntuación total en el test utilizada para agrupar a los sujetos en categorías es calculada incluyendo tanto los ítems con DIF como sin DIF. El nivel de habilidad estimado a los sujetos está, de esta manera, distorsionado. Los resultados obtenidos confirman esta hipótesis (véase la Figura 2). La aplicación del procedimiento MH bietápico frente a su aplicación en un primer y único análisis tiene efectos positivos análogos a los del procedimiento iterativo logit frente a los modelos loglinea-

les. En primer lugar, una severa reducción en el número de falsos positivos (reducción que es tanto más drástica cuanto mayor sea el tamaño de muestra). En segundo lugar, un incremento en el porcentaje de detecciones correctas, aunque en este caso las diferencias no sean tan marcadas (Figura 1). Estos resultados apoyan los hallazgos de Miller y Oshima (1992) en el sentido de que la principal ventaja de los métodos iterativos no es tanto el incremento en las tasas de detección, como la reducción en el número de falsos positivos.

También era esperable unas mayores tasas de identificaciones correctas en los ítems que presentaban una mayor diferencia en los índices de dificultad del ítem entre grupos ($d_1 = .1$, $d_2 = .19$, $d_3 = .34$, $d_4 = .43$). La mayoría de los estudios de simulación utilizan valores del área comprendidos entre 0.2 (DIF bajo) y 1.0 (magnitud de DIF elevada) (Swaminathan y Rogers, 1990; Narayanan y Swaminathan, 1994; Narayanan y Swaminathan, 1996). No es de extrañar, por consiguiente, las elevadas tasas de identificaciones correctas que presentan los ítems tercero (valor del área de 1.25) y cuarto (valor del área de 1.875) en todos los tamaños de muestra.

Respecto de la comparación del procedimiento MH bietápico y el método iterativo logit, cabe resaltar que el procedimiento MH bietápico se muestra más eficaz en la detección de los ítems con DIF que el método iterativo logit, con independencia del tamaño de muestra. Por el contrario, el método iterativo logit presenta, en general, un menor número de falsos positivos que el procedimiento MH bietápico. Se hipotetizó que este último resultado puede deberse a que el método iterativo logit elimina sucesivamente a los ítems que presentan DIF en los análisis anteriores, obteniendo de esta forma una mejor estimación del nivel de habilidad de los examinados. La aplicación iterativa del procedimiento MH corroboró esto como puede observarse de forma gráfica en la Figura 3.

Las principales conclusiones que pueden extraerse de los resultados expuestos son:

1. Los procedimientos iterativos tienen mayor potencia de prueba que los correspondientes procedimientos no iterativos. Sin embargo, su mayor ventaja es la reducción del error tipo I que conllevan.
2. A la vista de los resultados obtenidos no parece muy aconsejable la utilización de los modelos loglineales en una primera y única aplicación frente al método iterativo logit, ya que este último presenta un porcentaje de identificaciones correctas mayor que el primero con tamaños de muestra de 500 o más examinados por grupo. Además el error tipo I nunca supera la tasa nominal, llegando a ser el porcentaje de falsos positivos un 44% menor para el método iterativo logit frente a los modelos loglineales cuando $N = 1{,}000$.
3. En general, el procedimiento MH bietápico es preferible al procedimiento logit iterativo en la detección de DIF uniforme bajo las condiciones simuladas. No obstante, se puede arguir que trabajando con datos empíricos no se sabe a priori si el DIF es uniforme o no uniforme, por lo que, en estas condiciones, quizá sea más pertinente utilizar procedimientos que nos permitan detectar el DIF no uniforme en caso de producirse (por ejemplo, el procedimiento logit iterativo o la regresión logística).
4. La aplicación iterativa del procedimiento MH produce una mejora tanto en la tasa de identificaciones correctas como, sobre todo, en la reducción en el número de falsos positivos, por lo que parece aconsejable esta forma de cálculo.

Referencias

Ferrando, P. J. (1994). Saturaciones factoriales e índices de discriminación en la teoría clásica del test y en la teoría de respuesta a los ítems. *Anuario de Psicología, 62*, 55-65.

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*, 300.

Fidalgo, A. M. (1995). Differential item functioning [recensión del libro *Differential item functioning*. En W. P. Holland, y H. Wainer, (Eds.), 1993. Hillsdale, NJ: LEA]. *Psicothema, 7*, 237-241.

Fidalgo, A. M. (1996a). Funcionamiento diferencial de los ítems. En J. Muñiz (Coord.), *Psicometría* (pp. 370-455). Madrid: Universitas.

Fidalgo, A. M. (1996b). *Funcionamiento diferencial de los ítems. Procedimiento Mantel-Haenszel y modelos loglineales*. Tesis Doctoral, Universidad de Oviedo.

Fidalgo, A. M. y Mellenbergh, G. J. (1995, Abril). *Evaluación del procedimiento Mantel-Haenszel frente al método logit iterativo en la detección del funcionamiento diferencial de los ítems uniforme y no uniforme*. Comunicación presentada al IV Symposium de Metodología de las Ciencias del Comportamiento, La Manga del Mar Menor, Murcia.

Fidalgo, A. M. , Mellenbergh, G. J. y Muñiz, J. (1997a). *A comparison of the Mantel -Haenszel procedure and the iterative logit metod for detecting nonuniform differential item functioning*. Enviado.

Fidalgo, A. M. , Mellenbergh, G. J. y Muñiz, J. (1997b). *Effects of computing the MH chi-square statistic in a single stage, in two-stage and iteratively for detecting item functioning*. Enviado.

Fidalgo, A. M. y Paz, M. D. (1995a). Modelos lineales logarítmicos y funcionamiento diferencial de los ítems. *Anuario de Psicología, 64*, 57-66.

Fidalgo, A. M. y Paz, M. D. (1995b, Abril). *Comparación del método logit iterativo frente a los modelos loglineales en la detección del funcionamiento diferencial de los ítems*. Comunicación presentada al IV Symposium de Metodología de las Ciencias del Comportamiento, La Manga del Mar Menor, Murcia.

Gómez, J. y Navas, M. J. (1996). Detección del sesgo mediante regresión logística: purificación paso a paso de la habilidad. *Psicológica, 17*, 397-411.

Hambleton, R. K. y Rogers, H. J. (1989). Detecting potentially biased test ítems: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2* (4), 313-334.

Hidalgo, M. D. (1995). *Evaluación del funcionamiento diferencial del ítem en ítems dicotómicos y politómicos: un estudio comparativo*. Tesis Doctoral no publicada, Universidad de Murcia.

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

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54*, 681- 697.

Kim, S. y Cohen, A. S. (1991). A comparison of two areas measures for detecting differential item functioning. *Applied Psychological Measurement, 15*, 269-278.

Kok, F. G. , Mellenbergh, G. J. y Van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement, 22*, 295-303.

Lim, R. G. y Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75*, 164-174.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, N J: LEA.

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

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105-107.

Miller, M. D. y Oshima, T. C. (1992). Effect of sample size, number of biased ítems, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16*, 381-388.

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems.* Madrid: Pirámide.

Narayanan, P. y Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315-328.

Narayanan, P. y Swaminathan, H. (1996). Identification of item that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257-274.

Pardo, A. y San Martín, R. (1994). *Análisis de datos en psicología II.* Madrid: Pirámide.

Raju, N. S. (1988). The area between two item characteristics curves. *Psychometrika, 53*, 495-502.

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

Reynols, H. T. (1977). *The analysis of cross-classifications.* New York: Free Press.

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

Van der Flier, H. , Mellenbergh, G. J. , Adèr, H. J. y Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement, 21*, 131-145.

Aceptado el 4 de noviembre de 1997