Psicológica (1999) 20, 227-242.

Aplicación en una etapa, dos etapas e iterativamente de los estadísticos Mantel-Haenszel

Ángel M. Fidalgo*, Gideon J. Mellenbergh** y José Muñiz*

* Universidad de Oviedo

** Universidad de Amsterdam

En este estudio de simulación se examina el efecto que tres formas diferentes de aplicar el procedimiento Mantel-Haenszel - en una etapa, en dos etapas e iterativamente- tienen sobre los estadísticos: a) ji-cuadrado Mantel-Haenszel (χ^2_{MH}), y b) el cociente de razones común Mantel-Haenszel ($\hat{\alpha}_{MH}$). Los datos fueron simulados bajo dos tamaños de muestra (200 y 1,000 examinados por grupo), dos distribuciones de la habilidad (igual y diferente distribución de la habilidad entre el grupo focal y el de referencia), y dos porcentajes de items sesgados en el test (10% y 20%). El principal resultado del estudio es que el procedimiento bietápico y el iterativo siempre deben preferirse al de una sola etapa, al ser más robustos y presentar una adecuada potencia de prueba, además de ofrecer mejores estimaciones del parámetro α. Además, los resultados señalan la importancia de utilizar conjuntamente en la detección del funcionamiento diferencial de los items, tanto el estadístico $\hat{\alpha}_{MH}$ como el estadístico χ^2_{MH}, sobre todo con tamaños de muestra pequeños.

Palabras clave: funcionamiento diferencial de los items (DIF), cociente de razones común Mantel-Haenszel, ji-cuadrado Mantel-Haenszel, detección bietápica del DIF, detección iterativa del DIF.

A nadie puede extrañarle que garantizar la imparcialidad de los tests estandarizados haya llegado a ser uno de los núcleos centrales de las más recientes investigaciones psicométricas, dado el amplio uso que de los mismos se hace en los procesos de selección, promoción y certificación en los ámbitos educativos y profesionales. La abundancia de procedimientos para detectar qué items de un test están sesgados contra algún grupo en particular, o dicho de otra forma, funcionan diferencialmente[1], viene a corroborar de forma

* Correspondencia dirigirla a Angel. M. Fidalgo. Facultad de Psicología. Universidad de Oviedo. Plaza de Feijóo, s/n. 33003 Oviedo. Telf. 98 510 41 67. Fax 98 510 41 41. E-mail: fidalgo@pinon.ccu.uniovi.es

[1] A lo largo del artículo se utilizará, para evitar reiteraciones y farragosidad en la lectura, el término sesgo como sinónimo de funcionamiento diferencial. Todo lo dicho, por lo tanto,

empírica y efectiva la aseveración anterior. Hay muchos, pero no todos son iguales. Un requisito fundamental que deben cumplir dichos procedimientos estadísticos es que no confundan las diferencias reales entre los grupos en la habilidad medida, lo que técnicamente se denomina impacto, con las diferencias provocadas por un funcionamiento diferencial de los items. Por supuesto, el método más popular para detectar el funcionamiento diferencial de los items (differential item functioning, DIF), el procedimiento Mantel-Haenszel (MH) propuesto por Holland y Thayer (1988), cumple el requisito anteriormente expuesto. Las razones del amplio uso de este procedimiento son su sencillez, bajo costo computacional, buenos resultados, y el hecho de proporcionar tanto un estimador de la magnitud del sesgo presente en el ítem (el cociente de razones común Mantel-Haenszel, $\hat{\alpha}_{MH}$), como un test de significación estadística (el estadístico ji-cuadrado Mantel-Haenszel, χ^2_{MH}). El hecho de que el procedimiento MH sea uno de los más utilizados ha generado gran cantidad de estudios de simulación para determinar cómo se ve afectado por variables como el tamaño de muestra, el porcentaje de items sesgados, o la presencia de iguales o diferentes distribuciones entre grupos en la variable que mide el test, entre otras (Allen y Donoghue, 1996; Donoghue, Holland y Thayer,1993; Ferreres, Fidalgo y Muñiz, 1999; Fidalgo, Mellenbergh y Muñiz, 1998,1999; Miller y Oshima, 1992; Narayanan y Swaminathan, 1994, 1996; Parshall y Miller, 1995; Rogers y Swaminathan, 1993; Roussos y Stout, 1996; Uttaro y Millsap, 1994). La mayoría de estos estudios de simulación se han centrado en el estadístico χ^2_{MH}, siendo muy pocos los que han investigado las propiedades del estimador de la magnitud del DIF, el estadístico $\hat{\alpha}_{MH}$ (Allen y Donoghue, 1996; Donoghue, Holland y Thayer,1993; Roussos y Stout, 1996; Uttaro y Millsap, 1994). Sin embargo, es importante conocer cómo se ve afectado el estadístico $\hat{\alpha}_{MH}$ por variables como el tamaño de muestra, la longitud del test o cualesquiera otras, ya que la clasificación de un ítem como sesgado o insesgado no debe depender sólo del valor obtenido en el estadístico χ^2_{MH}. Así por ejemplo, puede ocurrir que bajo determinadas condiciones la tasa de error de Tipo I asociada con χ^2_{MH} sea bastante elevada, esto es, que concluyamos erróneamente que un gran número de items insesgados están sesgados. En estas condiciones es importante saber si los valores obtenidos por el estimador del DIF son elevados o no. Una tasa de error de Tipo I por encima de la esperada teóricamente no será un serio problema si los valores del estadístico $\hat{\alpha}_{MH}$ están próximos a 1, indicándonos ausencia de DIF. También puede ocurrir lo contrario, que en situaciones en que la potencia de prueba del estadístico χ^2_{MH} para detectar los items con DIF sea baja, tengamos unos valores en el estadístico $\hat{\alpha}_{MH}$ lo suficientemente alejados de 1 para indicarnos que el ítem está sesgado. Algunos de los resultados obtenidos en los estudios de simulación antes citados sugieren que

---

se referirá al funcionamiento diferencial de los items. Para una discusión pormenorizada de los conceptos de funcionamiento diferencial y sesgo de los items remitimos al lector a los siguientes textos de Fidalgo (1995 y 1996b).

el procedimiento MH puede indicar falsamente DIF (en términos tanto de $\hat{\alpha}_{MH}$ como de χ^2_{MH}) en tests de reducido tamaño (20 items o menos), cuando los datos no se ajustan al modelo de Rasch y existe impacto (Uttaro y Millsap, 1994). De otra parte, Donoghue, Holland y Thayer (1993) en un estudio en el que se manipularon los parámetros a y b del ítem bajo estudio, la inclusión o exclusión del ítem estudiado en la variable de agrupamiento, la longitud del test, el número de items sesgados en el test y la magnitud del DIF, encontraron que los factores que más influían sobre $\hat{\alpha}_{MH}$ fueron la inclusión del item bajo estudio en la variable de agrupamiento, su índice de dificultad, y la cantidad de DIF que presentaba. Una limitación que presentan los estudios de Donoghue, Holland y Thayer (1993) y Uttaro y Millsap (1994) es el elevado tamaño de muestra empleado en cada uno de los grupos (500 sujetos en el grupo de menor tamaño). En el trabajo de Roussos y Stout (1996) sí se manipula el tamaño de muestra, encontrando que el incremento en la potencia de prueba del MH con el tamaño de muestra no se debe al correspondiente incremento en $\hat{\alpha}_{MH}$; resultados similares fueron obtenidos por Fidalgo (1996a). Sin embargo, Roussos y Stout (1996) no estudiaron el efecto que el porcentaje de items con DIF en el test pudiera tener sobre el estadístico $\hat{\alpha}_{MH}$. Por contra de los anteriores, en el presente estudio se manipularán ambas variables: el tamaño de muestra y el porcentaje de items con DIF.

De otra parte, numerosas investigaciones recomiendan el uso de procedimientos que refinen las estimaciones de la habilidad de los sujetos en la variable medida, eliminando aquellos items que sean encontrados sesgados en los análisis iniciales. Diversos procedimientos iterativos han sido propuestos, algunos basados en el análisis de tablas de contingencia como el método iterativo logit (Van der Flier, Mellenbergh, Adèr y Wijn, 1984), la regresión logística (Gómez y Navas, 1996) o la utilización en dos etapas del procedimiento MH (Holland y Thayer, 1988), y otros que usan modelos de teoría de respuesta a los items (TRI) (Candell y Drasgow, 1988; Lautenschlager, Flaherty y Park, 1994; Lord, 1980; Miller y Oshima, 1992). En general, estos estudios indican que los procedimientos iterativos obtienen mejores resultados en la detección de DIF que cuando los correspondientes métodos son aplicados en un inicial y único análisis. En los citados estudios dicha mejora fue siempre operacionalizada en términos de potencia de prueba y tasa de error de Tipo I, es decir, en el caso del procedimiento MH, centrándose en el análisis del estadístico χ^2_{MH} (Fidalgo, 1996a; Fidalgo, Mellenbergh y Muñiz, 1998, 1999; Miller y Oshima, 1992). Por contra, hasta la fecha no hay ningún estudio que señale cómo se ve afectado el estadístico $\hat{\alpha}_{MH}$ por la aplicación iterativa del procedimiento MH. Este será otro de los objetivos de la presente investigación.

En este contexto, nuestra principal meta ha sido examinar, usando datos simulados, la influencia que el tamaño de muestra, la distribución de la habilidad y el porcentaje de items con DIF tienen sobre χ^2_{MH} y $\hat{\alpha}_{MH}$, cuando el procedimiento MH se aplica en un único e inicial análisis, en dos etapas e iterativamente.

MÉTODO

Generación de los datos. Con el fin de crear condiciones que fuesen representativas de las encontradas en situaciones reales, los parámetros utilizados fueron obtenidos a partir de la aplicación en 1985 de los 70 items que componen el *Graduate Management Admission Test* (Kingston, Leary y Wightman, 1988). Clauser, Mazor y Hambleton (1994) ajustaron un modelo logístico de tres parámetros a estos datos, obteniendo los parámetros que se presentan en la Tabla 1.

Tabla 1. Parámetros de los items utilizados para la generación de los datos.

	Parámetros			Parámetros			Parámetros	
Item nº	b	a	Item nº	b	a	Item nº	b	a
1	.44	.64	26	.06	.44	51	.51	.55
2	-.42	.75	27	-2.41	.27	52	.80	.53
3	1.39	1.04	28	3.47	.67	53	.16	.91
4	-1.17	.47	29	1.43	1.10	54	-.12	.84
5	.28	.61	30	-1.23	.59	55	1.29	.10
6	-1.47	.32	31	1.40	.95	56	.07	.69
7	.37	.72	32	2.27	.70	57	-.47	.44
8	.97	.76	33	.26	.71	58	-.45	.53
9	-1.11	.15	34	2.26	1.30	59	1.61	.69
10	-1.13	.28	35	.99	.22	60	-2.27	.24
11	.22	1.00	36	-1.73	.37	61	1.24	.61
12	-1.07	.30	37	.64	.49	62	1.41	.75
13	.03	.93	38	-1.12	.57	63	.09	.85
14	-.18	.83	39	-.91	.43	64	.59	.96
15	-1.61	.54	40	.33	1.05	65	.75	.59
16	-.91	.40	41	-2.09	.34	66	-.78	.63
17	.12	.34	42	.55	1.27	67	.85	1.16
18	-1.20	.38	43	-.19	.33	68	-1.70	.43
19	.94	.73	44	1.74	1.21	69	-.23	.43
20	1.22	.42	45	.23	.40	70	.24	.79
21	-2.25	.51	46	-1.05	.58			
22	-1.30	.61	47	.73	1.30			
23	1.23	.82	48	.19	.33			
24	-2.08	.46	49	1.15	.39			
25	.88	.47	50	.51	.78			

Nota. El parámetro c de todos los items es igual a .2

Se ha utilizado un modelo logístico de TRI de 3 parámetros para generar los datos. La habilidad en el grupo de referencia sigue una distribución normal con media 0 y desviación típica de 1 N(0,1). Se simularon dos grupos focales: el primero con la misma distribución de habilidad que el grupo de referencia, y el segundo con una media 1 desviación típica por

debajo de la media del grupo de referencia N(-1, 1). Conociendo la habilidad de cada examinado y los valores de los parámetros de los ítems, la probabilidad de una respuesta correcta en un modelo logístico de tres parámetros viene dada por:

$$P_i(\theta) = c_i + \{[1- c_i] / [1+ \exp(1.7\, a_i (\theta - b_i))]\}$$

Cada examinado recibía una respuesta correcta (puntuación de 1) cuando un número extraído al azar de una distribución uniforme en el intervalo (0,1) era menor o igual que la probabilidad de una respuesta correcta, de otra forma el examinado recibía una puntuación de 0.

Diseño. Para estudiar el efecto que aplicar el procedimiento MH en una etapa (MH-1), en dos etapas (MH-2) e iterativamente (MH-I) tiene sobre los estadístico MH, 8 condiciones fueron simuladas. Estas condiciones se obtuvieron cruzando dos tamaños de muestra (200 y 1,000 examinados por grupo), dos niveles en la distribución de la habilidad del grupo focal [N(0,1) y N(-1, 1)], y dos niveles de porcentaje de items con DIF (10% y 20%). Se eligieron tamaños de muestra de 1,000 y 200 sujetos por grupo por representar la primera condición (1,000 sujetos) una situación óptima, y raramente conseguida, para la aplicación de los estadísticos estudiados, y la segunda condición (200 sujetos) por ser el tamaño mínimo de muestra para emplear con garantías el procedimiento MH según Mazor, Clauser y Hambleton (1992). Para estudiar el efecto que la proporción de items con DIF en el test tiene sobre el procedimiento MH se utilizó en una condición un porcentaje del 10%, por ser habitual encontrar entre un 10 y un 15% de items sesgados en test estandarizados de aptitudes (Clauser, 1993), empleándose en la otra condición un 20% de items con DIF para ponernos en la peor de las situaciones. En la condición de impacto se utilizó una diferencia entre los grupos de una desviación típica. Este valor es habitualmente empleado en los estudios de simulación por ser representativo de las diferencias encontradas con datos reales cuando los grupos difieren significativamente en la habilidad medida (Roussos y Stout, 1996). En cada una de las 2 x 2 x 2 = 8 condiciones del diseño se generaron 200 conjuntos de datos, cada uno de los cuales fue analizado utilizando los procedimientos MH-1, MH-2 y MH-I.

El DIF se ha introducido variando el parámetro b en el grupo focal. La magnitud del DIF fue cuantificada en términos del área entre las curvas características del ítem (CCI) de cada uno de los grupos, utilizando la fórmula de Raju (1988) nº 7 para modelos de tres parámetros. Se generó una cantidad de DIF moderada (una diferencia de 0.50 entre las de CCI de cada uno de los grupos). Los items con DIF fueron más difíciles para el grupo focal que para el de referencia: $b_F = b_R + [0.50 / (1 - c)]$. En la condición del 20% de DIF, los parámetros b en el grupo focal de los items 57 a 63 serán cambiados para presentar la magnitud de DIF establecida (0.5). En la condición del 10% de DIF, el valor del parámetro b de los items 57 a 63 en el grupo focal será el mismo que en el grupo de referencia, presentando sólo DIF los items del 64 al

70. El valor del parámetro b en los items 64 a 70 será cambiado para presentar DIF moderado (0.5), tanto en la condición del 10% como en la del 20%. Los items que se analizarán serán siempre los siete últimos (items 64 a 70). El valor del parámetro b en los items con DIF se presenta en la Tabla 2.

Tabla 2. Parámetros b en el grupo focal utilizados para generar un grado moderado de DIF (0.50).

Item nº	b
57	.1550
58	.1750
59	2.2350
60	-2.1450
61	1.8650
62	2.0350
63	.1550
64	1.2150
65	1.3750
66	-.1550
67	1.4750
68	-1.0750
69	.3950
70	.8650

Análisis. El programa MHDIF (Fidalgo, 1994) fue utilizado para analizar cada base de datos. Dicho programa implementa el procedimiento bietápico propuesto por Holland y Thayer (1988). Construye k + 1 categorías (siendo k el número de items en el test) y calcula el estadístico χ^2_{MH}. Elimina los items con valores significativos en el estadístico χ^2_{MH}, y recalcula de nuevo los estadísticos MH (χ^2_{MH} y $\hat{\alpha}_{MH}$) usando la puntuación total del sujeto en los items 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 previo. Además se elaboró un programa que calculase el procedimiento MH de forma iterativa. El citado programa procede de la siguiente forma: Construye k + 1 categorías (siendo k el número de items en el test) y calcula el estadístico χ^2_{MH}. Elimina los items con valores significativos en el estadístico χ^2_{MH}, y recalcula de nuevo los estadísticos MH usando la puntuación total del sujeto en los items restantes como criterio de bloqueo. Así procede iterativamente hasta que (a) los items identificados con DIF son los mismos que los identificados en la iteración anterior, o (b) se llega al número máximo de iteraciones fijado (15). Como el programa MHDIF, cuando un ítem está siendo investigado se incluye en el criterio de agrupamiento aunque haya presentado DIF en el análisis previo.

RESULTADOS

La calidad de un contraste estadístico habitualmente es valorada en términos de potencia de prueba y robustez. Se dice que un estadístico es robusto si su probabilidad empírica ($\hat{\Pi}$) de cometer errores de Tipo I es aproximadamente igual al nivel de significación utilizado en los análisis. Bradley (1978) propuso un criterio liberal y otro conservador para determinar en qué medida un test estadístico se alejaba de forma más o menos extrema de sus valores óptimos. Un test satisface el criterio liberal de Bradley si $0.5\alpha \leq \Pi \leq 1.5\alpha$, y el criterio conservador si $0.9\alpha \leq \Pi \leq 1.1\alpha$. En cada una de las condiciones del diseño se tomó como estimador de la tasa de error de Tipo I ($\hat{\Pi}$) la proporción de falsos positivos ocurrida en los items 1 a 56 (aquellos que no están sesgados ni en la condición de 10% de DIF, ni en la del 20%) a lo largo de las 200 replicaciones. Estas tasas empíricas cumplirán el criterio liberal si $0.025 \leq \hat{\Pi} \leq 0.075$ y el conservador si $0.045 \leq \hat{\Pi} \leq 0.055$.

La potencia de prueba de un contraste estadístico , o simplemente potencia, es la probabilidad con que rechaza correctamente la hipótesis nula. Se tomó como estimador de la potencia del estadístico χ^2_{MH} para detectar cada uno de los 7 items con DIF analizados (items 64 a 70), la proporción de veces que cada ítem fue detectado como sesgado a lo largo de las 200 replicaciones.

Items con DIF (items 64 a 70)

Los resultados muestran, a lo largo de las 8 condiciones, escasas diferencias entre los valores del estadístico $\hat{\alpha}_{MH}$ en función de la forma de cálculo (MH-1, MH-2, MH-I). La potencia de prueba tampoco varió sustancialmente en función de la forma de cálculo. Las diferencias, eso sí, cuando las hubo, siempre fueron en beneficio de la aplicación bietápica e iterativa del procedimiento MH. Para ilustrar este punto, en la Figura 1 podemos observar como la potencia de prueba, promediada a lo largo de todas las condiciones y de todos los items sesgados analizados, es muy similar entre las diversas formas de aplicar el MH.

En la Figura 2 para poder visualizar de una forma directa y rápida el efecto que las variables manipuladas tienen sobre los estadísticos MH y, a la vez, en qué medida el comportamiento del estadístico $\hat{\alpha}_{MH}$ es simétrico al del estadístico χ^2_{MH}, se presentan una serie de gráficos en los que aparecen la potencia de prueba del estadístico χ^2_{MH} junto con el valor medio obtenido en $\hat{\alpha}_{MH}$ por los items analizados a lo largo de las 200 replicaciones. No se han presentado resultados separados para cada método de análisis porque, como se ha señalado antes (véase la Figura 1), las diferencias son mínimas. Vemos también como el comportamiento del estadístico $\hat{\alpha}_{MH}$ es simétrico al de χ^2_{MH}, es decir, que aquellos items que son detectados más probablemente, también

son aquellos en los que la estimación de la magnitud del DIF es mayor. En la Figura 1 también se observa claramente como el tipo de ítem influye en la probabilidad de ser detectado como sesgado así como en el valor obtenido en $\hat{\alpha}_{MH}$. Los items con mayor probabilidad de ser detectados son los items más discriminativos (items 70, 66 y 64), siendo los peor detectados aquellos que tienen un parámetro a más bajo (items 68, 69 y 65). Además la influencia del parámetro a parece estar modulada por el parámetro de dificultad del ítem, de forma que los items con el parámetro b más elevado (items más difíciles) son peor identificados independientemente de ser muy discriminativos (véase como ejemplo el ítem 67).

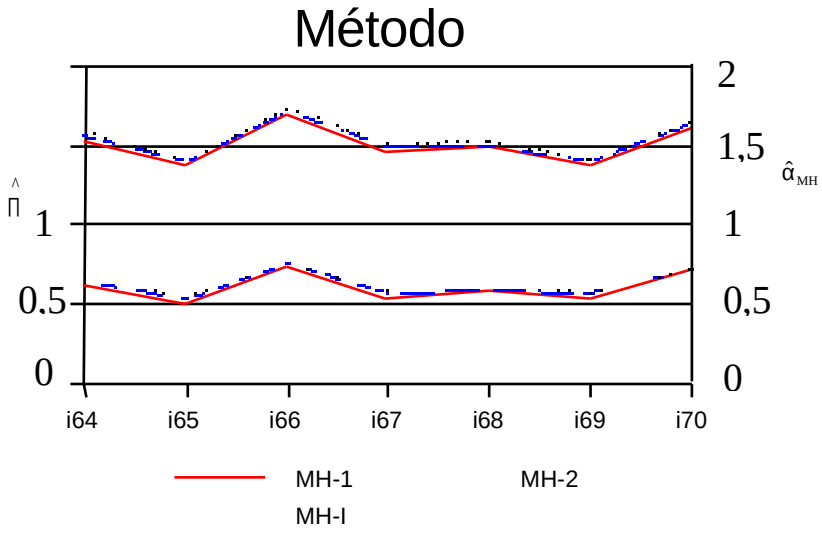

Figura 1. Potencia de prueba de χ^2_{MH} (lineas inferiores) y valores del estadístico $\hat{\alpha}_{MH}$ (lineas superiores) para la aplicación en una sola etapa (MH-1), en dos (MH-2) e iterativamente (MH-I) del procedimiento MH para los items con DIF bajo estudio (items 64 a 70).

El tamaño de muestra, como era esperable, tiene un fuerte efecto sobre la potencia de prueba del estadístico χ^2_{MH} (véase la Figura 2). Con N = 1,000, la proporción de detecciones varía desde 0.83 para el ítem 65 hasta 1.0 para el ítem 66, mientras que los valores mínimos y máximos con N = 200 fueron 0.21 (ítem 65) y 0.48 (ítem 66). Sin embargo, el tamaño de muestra no influye prácticamente sobre $\hat{\alpha}_{MH}$ (para N = 200 su valor promedio a lo largo de los items 56 a 70 fue de 1.53, y para N =1,000 de 1.51).

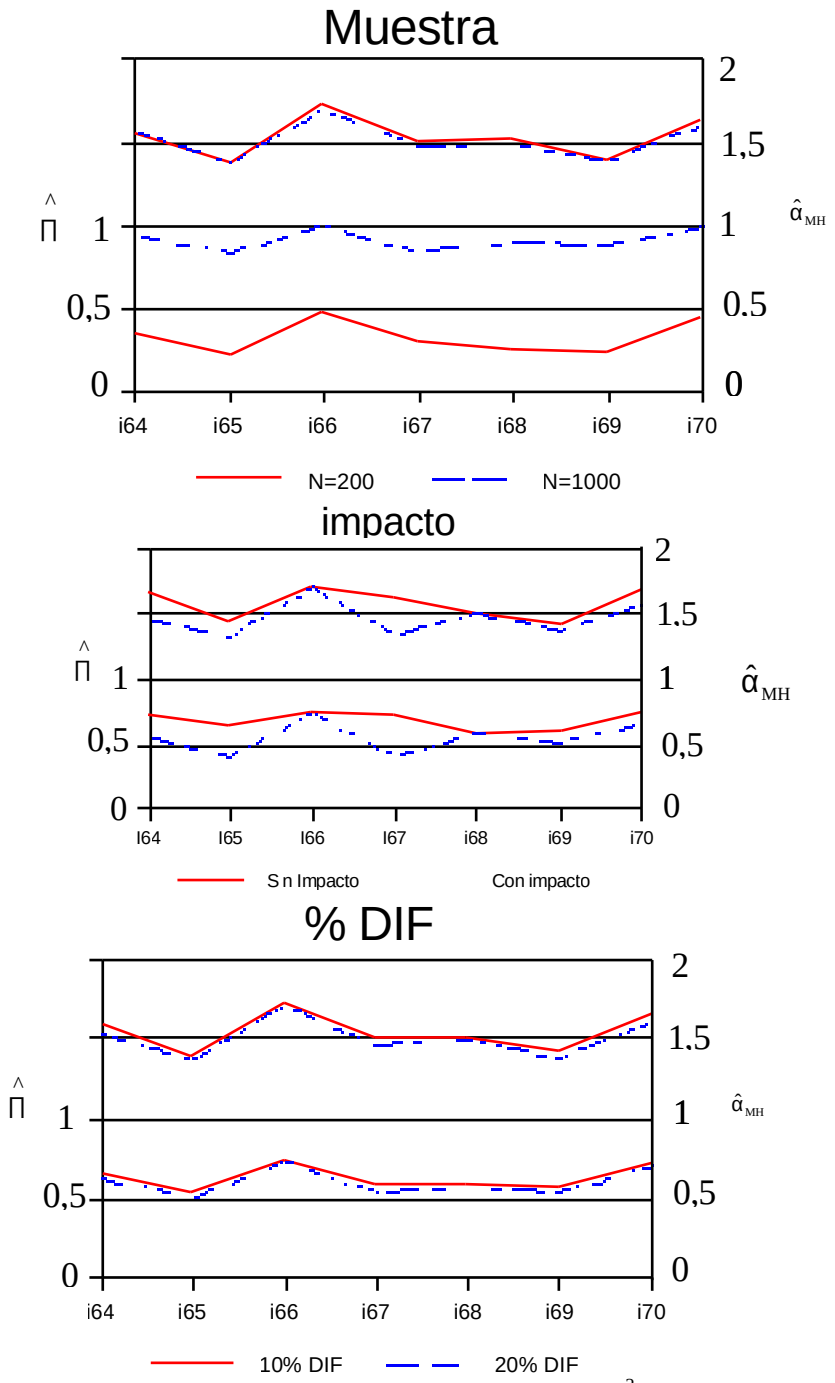

Figura 2. Gráficos con la potencia de prueba de χ^2_{MH} (lineas inferiores) y valores del estadístico $\hat{\alpha}_{MH}$ (lineas superiores) en cada una de las variables manipuladas para los items con DIF bajo estudio (items 64 a 70).

Como podemos apreciar en el gráfico que corresponde a la diferencia en la distribución de la habilidad entre los grupos o impacto en la Figura 2, el efecto de la diferencia en la distribución de la habilidad es una disminución tanto de la potencia de χ^2_{MH}, como de la magnitud de $\hat{\alpha}_{MH}$. Sin embargo, aquí, y a diferencia del resto de las variables consideradas, el efecto de los niveles de la variable impacto (igual distribución/ diferente distribución) sobre la potencia de χ^2_{MH} y la magnitud de $\hat{\alpha}_{MH}$ se ven altamente modulados por el tipo de ítem. Así los items que se ven menos afectados por la presencia o no de impacto son los que tienen un parámetro de dificultad más bajo (ítem 68 y 66), y, contrariamente, los que presentan un efecto diferencial mayor son aquellos cuyo parámetro b es mayor (items 67 y 65).

El porcentaje de items sesgados en el test tiene un efecto muy escaso sobre los estadísticos MH. Además, el comportamiento de ambos estadísticos, como puede observarse en la correspondiente gráfica de la Figura 2, es absolutamente simétrico.

Items sin DIF (items 1 a 56)

En la Tabla 3 se muestra el valor promedio de la proporción de falsos positivos para cada VI a lo largo de las 200 replicaciones. Estas proporciones son una estimación de la tasa de error de Tipo I del estadístico χ^2_{MH}, utilizando un nivel de significación de 0.05. En la Tabla 4 podemos observar el valor promedio obtenido a lo largo de las 200 replicaciones en el estadístico $\hat{\alpha}_{MH}$.

Tabla 3. Tasa de error de Tipo I para los factores manipulados en cada uno de los métodos de análisis

	Tipo de purificación		
Variables	MH-1	MH-2	MH-I
Muestra			
200	0.04*	0.04*	0.04*
1,000	0.09**	0.06*	0.05
Distribución habilidad			
Igual	0.07*	0.05	0.04*
Diferente	0.06*	0.05	0.05
% items con DIF			
10%	0.05	0.05	0.05
20%	0.07*	0.05	0.05

Nota: La presencia de un asterisco indica que la tasa de error de Tipo I estimada no cumple el criterio conservador de Bradley; dos asteriscos indican que tampoco cumple el criterio liberal.

Como podemos observar en la Tabla 3, el procedimiento MH aplicado en una sola etapa sólo cumple el criterio conservador de Bradley en una condición: cuando el porcentaje de items sesgados en el tests es del 10%. Además, es el único que presenta una tasa de error de Tipo I fuera de las bandas de confianza establecidas por el criterio liberal de Bradley. En efecto, cuando N = 1,000, presenta una tasa de error de Tipo I de 0.09, cuando el nivel nominal es el 0.05. El procedimiento bietápico, presenta una tasa de error de Tipo I por debajo de la fijada en el criterio conservador cuando N = 200 (0,04), y por encima (0,06) cuando N = 1,000. Por su parte, el procedimiento iterativo se muestra conservador (una estimación de 0.04) cuando tenemos un tamaño de muestra de 200 examinados, y también cuando la distribución de la habilidad es igual entre los grupos.

Tabla 4. Media de $\hat{\alpha}_{MH}$ en los items sin DIF (items 1 a 56) para los factores manipulados en cada uno de los métodos de análisis

	Tipo de purificación		
Variables	MH-1	MH-2	MH-I
Muestra			
200	0.98	0.99	1.00
1,000	0.95	0.99	1.00
Distribución habilidad			
Igual	0.95	0.99	1.00
Diferente	0.97	0.99	1.00
% items con DIF			
10%	0.98	1.00	1.01
20%	0.95	0.98	0.99

Respecto de $\hat{\alpha}_{MH}$ en los items sin DIF, en primer lugar debemos de hacer notar la escasa diferencia existente entre los diversos niveles de las VIs. La mayor diferencia se da entre las condiciones de 10% de items sesgados en el test y 20% de items sesgados (véase la Tabla 4). Como cabria esperar, cuanto mayor es el porcentaje de items sesgados más alejado se encuentra el valor promedio de $\hat{\alpha}_{MH}$ de 1. En los items sin DIF la aplicación del MH en una sola etapa, frente al procedimiento bietápico e iterativo, proporciona una infraestimación del parámetro α, aunque las diferencias entre los tres procedimientos, como puede observarse en la tabla 4, son mínimas.

DISCUSIÓN

Los resultados presentados muestran como tanto en los items sesgados como insesgados, las peores estimaciones del parámetro α se obtuvieron siempre cuando el procedimiento MH se aplicó en una etapa. Así por ejemplo, el procedimiento MH-1 fue el que proporcionó, para los items sesgados, menores estimaciones de la magnitud del DIF (un valor promedio de 1.49), seguido del bietápico (1.52) y del iterativo (1.54). También fue el que proporcionó, para los items insesgados, unas estimaciones del parámetro α más alejadas de 1 (el valor que se debe obtener en el caso de que no exista DIF). El efecto que aplicar el procedimiento MH en una etapa, en dos o iterativamente tiene sobre $\hat{\alpha}_{MH}$, como puede apreciarse en los promedios citados, es escaso. Lo mismo acontece respecto de la potencia de prueba (véase la Figura 1). El promedio de la potencia de prueba a lo largo de todas las condiciones fue para MH-1 de 0.59, de 0.62 para MH-2, y de 0.63 para MH-I. Sin embargo, si existen diferencias notables en la tasa de error de Tipo I entre el procedimiento inicial y los otros. Así por ejemplo, cuando el tamaño de muestra es de 1,000 examinados por grupo, mientras la tasa de error de Tipo I para el procedimiento iterativo fue de 0.05 y para el bietápico de 0.06, se elevó hasta 0.09 cuando el MH se aplicó una sola vez. Estos resultados se deben, sin duda, a que cuando el MH se aplica una única vez está utilizando los items sesgados en el cómputo de la puntuación en el test que servirá para establecer las comparaciones entre examinados de igual capacidad. Por otra parte, el procedimiento bietápico al eliminar de la puntuación en el test los items que aparecen con DIF en el primer análisis, obtiene mejores estimaciones del nivel de habilidad y mejora los resultados. Esta purificación de la habilidad es aún mayor cuando el procedimiento MH se aplica iterativamente.

Un resultado menos esperable que el anterior, es que el efecto que el impacto tiene sobre $\hat{\alpha}_{MH}$ y χ^2_{MH} en los items sesgados está altamente modulado por las características del ítem. Como se señaló, los items que se ven menos afectados por la presencia o no de impacto son los que tienen un parámetro de dificultad más bajo, mientras que los que presentan un parámetro $\underline{b}$ mayor, tienen una probabilidad mucho más elevada de no ser detectados, cuando existe impacto, que aquellos que son más fáciles. En el resto de las variables manipuladas, tamaño de muestra y porcentaje de items con DIF, las características del ítem no tienen un efecto diferencial sobre la influencia que los distintos niveles de dichas variables tienen sobre la potencia de prueba y el valor de $\hat{\alpha}_{MH}$ (véase la Figura 2). Además y en consonancia con lo encontrado en otros estudios (Hambleton, Clauser, Mazor y Jones, 1993), se comprobó como los items con mayor probabilidad de ser detectados fueron los items más discriminativos y aquellos con valores en el parámetro $\underline{b}$ más bajos.

También es destacable el efecto que el tamaño de muestra tiene sobre los estadísticos MH. Como era esperable, la potencia de prueba del estadístico χ^2_{MH} aumenta conforme lo hace el tamaño de muestra. Para $\underline{N}$ = 200 la

potencia de prueba de χ^2_{MH} fue muy baja (un promedio de 0.32), presentando unos valores aceptables (un promedio de 0.91) cuando $\underline{N}$ = 1,000. Sin embargo, el tamaño de muestra no influye prácticamente sobre $\hat{\alpha}_{MH}$ (para $\underline{N}$ = 200 su valor promedio fue de 1.53, y para $\underline{N}$ =1,000 de 1.51). Estos resultados hacen que la aplicación del estadístico χ^2_{MH} en muestra pequeñas, en contra de lo señalado por algunos autores (Hills, 1989; Mazor, Clauser y Hambleton, 1992) sea muy cuestionable. Sin embargo, dado el escaso efecto que el tamaño de muestra parece tener sobre $\hat{\alpha}_{MH}$, puede que en estas situaciones, sea mucho más fiable guiarnos por los valores obtenidos en $\hat{\alpha}_{MH}$ para detectar a los items sesgados, que por un estadístico de contraste con una escasa potencia de prueba. Ni que decir tiene que aplicar χ^2_{MH} nunca esta de más, aunque conociendo sus limitaciones.

Las principales conclusiones que podemos extraer del presente estudio se resumen en las siguientes. Primera, que el procedimiento de una etapa no es suficientemente robusto, sobre todo con tamaños de muestra elevados, mientras que el procedimiento bietápico e iterativo fueron, en general, robustos, además de presentar un potencia de prueba mayor y mejores estimaciones del parámetro α. Por lo tanto, se recomienda aplicar el MH en dos etapas o iterativamente. Señalar además que el escaso beneficio que supone el procedimiento iterativo frente al bietápico para las condiciones simuladas, es posible que no compense su mayor coste computacional. Segunda, en la detección del DIF con tamaños de muestra pequeños ($\underline{N}$ = 200) es necesario tomar en consideración a los resultados tanto de $\hat{\alpha}_{MH}$ como de χ^2_{MH} , dando prioridad a los valores obtenidos en el estadístico $\hat{\alpha}_{MH}$. Tercera, sería pertinente la realización de estudios que investiguen de forma conjunta y sistemática las relaciones existentes entre los estadísticos $\hat{\alpha}_{MH}$ y χ^2_{MH}, prestando especial consideración las variables que no han sido manipuladas de forma activa en la presente investigación, como el efecto que puedan tener los parámetros de los items investigados. También sería deseable, en un intento de aumentar la validez externa de los estudios de simulación, utilizar como parámetros de los items en la generación de los datos, estimaciones obtenidas en la calibración de tests aplicados en muestras españolas (Ferreres, 1998; Ferreres, González y Gómez, 1999), al modo en que se ha hecho en un estudio de simulación reciente (Ferreres, Fidalgo y Muñiz, 1999).

ABSTRACT

Computing single-, two-stage, and iterative Mantel-Haenszel statistics. This simulation study examined the effects of three different Mantel-Haenszel (MH) computing procedures (single-stage, two-stage and iterative procedures) on the Mantel-Haenszel statistics: a) MH chi-square (χ^2_{MH}), and b) MH common odds ratio estimator ($\hat{\alpha}_{MH}$). Data were simulated under two sample sizes (200 and 1,000 examinees per group), two ability distributions (equal and unequal ability distribution of focal and reference groups), and two percentages of DIF items in the test (10% and 20%).The main result of this study is that the two-stage and iterative MH-procedures must be preferred above the single-stage procedure because of robustness, overall higher power, and better estimates of α. Moreover, the results show the importance of using both MH statistics, for detecting differential item functioning when sample sizes are small.

Key words: differential item functioning (DIF), iterative DIF-detection, Mantel-Haenszel common of ratio estimator, Mantel-Haenszel chi-square statistic, two-stage DIF-detection.

REFERENCIAS

Allen, N. L. y Donoghue, J. R. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement, 33*, 231-251.

Bradley, J.V. (1978). Robustness? *The British Journal of Mathematical & Statistical Psychology, 31*, 144-152.

Candell, G. L. y Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*, 253-260.

Clauser, B.E. (1993). Factors influencing the performance of the Mantel-Haenszel procedure in identifying differential item functioning. Tesis doctoral, Universidad de Massachussetts at Amherts. *Dissertation Abstracts International, 54*, 493.

Clauser, B., Mazor, K. M. y Hambleton, R. K. (1994). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational Measurement, 31*, 67-78.

Donoghue, J. R. , Holland, P. W. y Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential Item

Functioning. En W. P. Holland y H. Wainer (Eds.), *Differential Item Functioning* (pp.137-166). Hillsdale, NJ: LEA.

Ferreres, D. (1998). *Funcionamiento diferencial de los items 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., Fidalgo, A.M., y Muñiz, J. (1999, Septiembre). *Detección del funcionamiento diferencial no uniforme: comparación de los métodos Mantel-Haenszel y regresión logística*. Comunicación presentada al VI Congreso de Metodología de las Ciencias Sociales y de la Salud, Oviedo.

Ferreres, D., González, V., y Gómez, J. (1999, Septiembre). *Comparación del estadístico Mantel-Haenszel y la regresión logística en el funcionamiento diferencial de los items en dos pruebas de aptitud intelectual en un contexto bilingüe*. Comunicación

presentada al VI Congreso de Metodología de las Ciencias Sociales y de la Salud, Oviedo.

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 items. Procedimiento Mantel-Haenszel y modelos loglineales*. Tesis Doctoral, Universidad de Oviedo.

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

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*, 219-228.

Fidalgo, A.M. y Mellenbergh, G.J. y Muñiz, J. (1999, Julio). *Robustness and power of single, two-stage, and iterative Mantel-Haenszel statistic procedures for DIF-detection*. Comunicación presentada a la ITC/IACCP Conference, Graz, Austria.

Gómez, J. y Navas, M. J. (1996). Detección del sesgo mediante regresión logística: purificación paso a paso de la habilidad. *Psicológica, 17*, 197-411.

Hambleton, R. K., Clauser, B. E., Mazor, K. M. & Jones, R. W. (1993). Advances in the detection of differencially functioning test items. *European Journal of Psychological Assessment, 9*, 1-18.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice, 8*, 5-11.

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: LEA.

Kingston, N., Leary, L. y Wightman, L. (1988). *An exploratory study of the applicability of item response theory methods to the Graduate Management Admissions Test* (GMAC Occasional Papers). Princenton, NJ: Graduate Management Admissions Council.

Lautenschlager, G. J., Flaherty, V. L. y Park, D. G. (1994). IRT differential item functioning: an examination of ability scale purifications. *Educational and Psychological Measurement, 54*, 21-31.

Lewis, C. (1993). A note on the value of including the studied item in the test score when analyzing test for DIF. En W.P. Holland y H. Wainer (Ed.), *Differential Item Functioning* (pp. 317-319). Hillsdale, NJ: LEA.

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

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

Mazor, K. M., Clauser, B. E. y Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*, 443-452

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

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.

Parshall, C. G. y Miller, T. R. (1995). Exact versus aymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small-sample conditions. *Journal of Educational Measurement, 32*, 302-316.

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

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*, 105-116.

Roussos, L. y Stout, W. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215-230.

Uttaro, T. y Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18*, 15-25.

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.

(Revisión aceptada: 2/11/99)