Anuario de Psicología
2007, vol. 38, nº 3, 431-442

Un meta-análisis del procedimiento Mantel-Haenszel en la detección del DIF en ítems dicotómicos*

Georgina Guilera
Juana Gómez-Benito
Universitat de Barcelona
Mª Dolores Hidalgo
Julio Sánchez-Meca
Universidad de Murcia

El objetivo del presente estudio consiste en describir el estado actual de la investigación sobre la eficacia del procedimiento Mantel-Haenszel (MH) como técnica de detección del Funcionamiento Diferencial del Ítem (Differential Item Functioning, DIF) *en ítems dicotómicos, y analizar la influencia de diversas variables sobre la potencia estadística y la tasa de Error Tipo I. Para ello, se ha llevado a cabo un estudio meta-analítico, siguiendo una estrategia de búsqueda de documentos en diversas bases de datos con la terminología* Differential item functioning, DIF, Mantel-Haenszel *y* MH*, hasta el año 2006. Por razones inherentes al objetivo del presente trabajo, se han incluido solamente los estudios con datos simulados. Para cada trabajo se han codificado las detecciones correctas de los ítems con DIF, así como los falsos positivos. Además se ha llevado a cabo un estudio detallado de las variables moderadoras que pueden afectar al funcionamiento de la técnica bajo estudio.*

Palabras clave: *Mantel-Haenszel, ítems dicotómicos, funcionamiento diferencial de ítem, meta-análisis.*

* Este estudio ha recibido el Premio AEMCCO para investigadores jóvenes 2007, otorgado en el marco del X Congreso de Metodología de las Ciencias Sociales y de la Salud, y ha sido financiado por los proyectos 2006FIC 00034 y 2005SGR00365 del Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya y SEJ2005-09144-C02-02/PSIC del Ministerio de Educación y Ciencia de España.
Correspondencia: Georgina Guilera. Departament de Metodologia de les Ciències del Comportament, Facultat de Psicologia, Universitat de Barcelona. Passeig de la Vall d'Hebron, 171. 08035 Barcelona (Spain).
Correo electrónico: gguilera@ub.edu

A meta-analysis of the Mantel-Haenszel procedure in the detection of DIF in dichotomous items

The study describes the present state of research into the effectiveness of the Mantel-Haenszel technique (MH) to detect Differential Item Functioning (DIF) in dichotomous items, and analyses the influence of diverse variables on the statistical power and the Type I error rate. A meta-analysis was performed, running searches in various databases for the terms Differencial Item Functioning, DIF, Mantel-Haenszel and MH*, up to 2006. Only studies with simulated data were included. For each study, the correct detections of items with DIF were coded, as well as the false positives. Moderator variables that may affect the procedure under study were also assessed.*

Key words: *Mantel-Haenszel, dichotomous items, differential item functioning, meta-analysis.*

Un determinado ítem presenta funcionamiento diferencial del ítem (*Differential Item Functioning*, DIF) si se comporta diferencialmente para individuos o grupos comparables, que difieren en lengua nativa, género, etnia, cultura, o cualquier otra variable que pueda constituir una fuente sistemática de variación ajena al rasgo medido por la prueba en cuestión, entendiendo por comparables aquellos grupos de sujetos que poseen el mismo nivel en la característica o rasgo medido por el test (Gómez e Hidalgo, 1997). En la terminología propia del DIF, se denomina grupo focal (F) al conjunto de individuos, generalmente minoritario, que representa el foco de interés del estudio y que normalmente es el grupo desaventajado, mientras que el grupo de referencia (R), generalmente mayoritario, se refiere a un grupo de sujetos estándar respecto al cual se compara el grupo focal.

A finales de los años 50, Mantel y Haenszel (1959) propusieron un nuevo método para el análisis de tablas de contingencia tridimensionales que permitía estudiar diferencias entre grupos comparables; posteriormente Holland (1985) y Holland y Thayer (1988) adaptaron este procedimiento para utilizarlo como técnica de detección del DIF. Esta técnica, ampliamente extendida y preferida por el *Educational Testing Service*, compara la ejecución de un ítem en el grupo de referencia y el grupo focal a través de distintos niveles de una determinada variable de equiparación (criterio o habilidad); en este sentido, se asume que en cada uno de estos niveles los individuos de uno y otro grupo son comparables, y si un ítem no presenta DIF lo ejecutarían por igual, presentando la misma probabilidad de acierto o error.

Con el propósito de comparar las probabilidades de acierto a un ítem, los datos del grupo de referencia y del grupo focal se distribuyen en tantas tablas de contingencia 2×2 como niveles en la habilidad de los sujetos; dichos niveles normalmente se diseñan a partir de la puntuación total de los sujetos en el test. La tabla de contingencia 2×2 para el ítem i en el nivel de habilidad j se muestra en la tabla 1.

La hipótesis nula de ausencia de DIF postula que la probabilidad de acertar el ítem bajo estudio en el nivel de habilidad j es la misma para el grupo de referencia

TABLA 1. TABLA DE CONTINGENCIA 2×2 EN EL NIVEL DE HABILIDAD *j*

	Puntuación en el ítem		
Grupo	*Acierto (1)*	*Error (0)*	*Total*
Referencia	A_j	B_j	$N_{R.j}$
Focal	C_j	D_j	$N_{F.j}$
Total	$N_{.1j}$	$N_{.0j}$	$N_{..j}$

que para el grupo focal, mientras que la hipótesis alternativa de presencia de DIF formula que esa probabilidad de acierto en el grupo de referencia equivale a un cociente de razones común, denominado α, multiplicado por la probabilidad de acierto en el grupo focal.

La expresión de α viene dada por:

$$\alpha_{MH} = \frac{\sum A_j D_j / N_{..j}}{\sum B_j C_j / N_{..j}},$$

pudiendo adoptar valores entre 0 e infinito.

Por tanto, la hipótesis nula de ausencia de DIF está representada por un alfa con valor 1, mientras que un valor mayor que 1 indica que el grupo de referencia presenta una probabilidad más elevada de acertar el ítem que el grupo focal y, del mismo modo, un valor menor que 1 indica que el grupo aventajado es el focal frente al de referencia.

Además el procedimiento MH ofrece un estadístico, asociado a una prueba de significación, que sigue una distribución χ^2 con un grado de libertad:

$$\chi^2_{MH} = \frac{\left[\left|\sum A_j - \sum E(A_j)\right| - 0.5\right]^2}{\sum Var(A_j)},$$

donde

$$E(A_j) = \frac{N_{R.j} N_{.1j}}{N_{..j}},$$

y

$$Var(A_j) = \frac{N_{R.j} N_{F.j} N_{.1j} N_{0.j}}{(N_{..j})^2 (N_{..j} - 1)}$$

Aunque el uso de esta técnica ha recibido diversas críticas (ver Mellenbergh, 1982), el procedimiento MH es conceptualmente sencillo, relativamente fácil de implementar, y ofrece una prueba de significación estadística, por lo que ha sido extensamente utilizado como método para la detección del DIF (Narayanan y Swaminathan, 1996) hasta el punto de que es la técnica adoptada por el *Educational Testing Service*, una organización de referencia en este área de cono-

cimiento. En este contexto y tras más de dos décadas de la aplicación del procedimiento MH como técnica de detección del DIF, son muchos los trabajos que se han dedicado a analizar el Error Tipo I y/o la potencia estadística de este procedimiento, cada uno de los cuales ha evaluado el funcionamiento de MH bajo unas condiciones determinadas mediante la simulación de datos. Sin embargo, no hay constancia de la existencia de ningún trabajo que pretenda integrar cuantitativamente los resultados de estas investigaciones posibilitando así una mayor generalización de los resultados del funcionamiento de la técnica MH.

Así, el objetivo del presente trabajo consiste en realizar un primer meta-análisis sobre la eficacia del procedimiento MH como técnica de detección del DIF en ítems dicotómicos con el propósito de conocer en más profundidad este procedimiento y proporcionar ayuda a los investigadores para aplicarlo en las condiciones más adecuadas; concretamente, en un primer estudio, se pretende determinar bajo qué condiciones es menor la tasa de Error Tipo I y, en un segundo estudio, bajo qué condiciones presenta una potencia estadística más elevada. Finalmente, se ofrece un estudio descriptivo donde se detallan las condiciones donde el Error Tipo I y la potencia del MH son adecuadas.

Método

Criterios de inclusión de estudios

Por razones inherentes al objetivo del presente trabajo, la búsqueda se limitó a estudios que: *a)* empleasen datos simulados; *b)* utilizasen la prueba MH como técnica de detección del DIF en ítems dicotómicos; *c)* evaluasen la tasa de Error Tipo I y/o la potencia de la técnica; *d)* utilizasen una prueba de significación para MH a la hora de determinar la tasa de detección a un nivel del 5%; *e)* no empleasen procedimientos de purificación de la variable de equiparación; y *f)* presentasen los datos necesarios para obtener los resultados (frecuencias, porcentajes o proporciones de detección). Los puntos *d)* y *e)* se incluyeron como criterios de selección de los trabajos, dado que en el uso de procedimientos meta-analíticos es importante mantener la independencia entre los resultados que se pretenden integrar.

Solamente se incluyeron los trabajos publicados en revistas científicas. Los trabajos que no cumplieron todos y cada uno de los criterios de inclusión, fueron retirados del estudio.

Identificación de los documentos

La localización de los estudios se realizó de forma electrónica en las bases de datos Medline, Eric, Psychology and Behavioral Sciences Collection, PsycInfo, Sciences Citation Index-Expanded, Social Sciences Citation Index y Arts and Humanities Citation Index en octubre del 2006 siguiendo la estrategia de búsqueda que a continuación se detalla:

(differential item functioning or DIF) and (Mantel-Haenszel or MH)

Recopilación y codificación de los datos

Además de los documentos identificados mediante la estrategia de búsqueda electrónica y con el objetivo de no omitir ningún artículo, se revisaron las referencias de cada uno de los artículos.

Las variables moderadoras que se tuvieron en cuenta fueron las propias de cualquier estudio de simulación de este tipo, donde se contemplan aquellas variables que pueden estar influyendo en la tasa de error y/o potencia del procedimiento MH. Concretamente se tuvieron en cuenta: el modelo de medida de simulación de los datos, el valor del parámetro de discriminación (a), el valor del parámetro de dificultad (b), el valor del parámetro de pseudo-azar (c), el número de ítems que contiene el test, la cantidad de impacto entre los grupos a comparar (focal y de referencia), el número de sujetos del grupo de referencia, el número de sujetos del grupo focal, el tipo de DIF (uniforme o no-uniforme), la cantidad de DIF medida como el área definida por la diferencia entre las curvas características de los ítems, y el tanto por ciento de ítems en el test que presentan DIF.

Se elaboró un libro de codificación donde se especificaron las variables a codificar, así como sus correspondientes categorías y, en los casos que fue necesario, se añadió información complementaria para garantizar un adecuado proceso de codificación. Se clasificó el parámetro *a* en bajo ($a \leq 0.5$), medio ($0.5 < a < 1.5$) y alto ($a \geq 1.5$); el parámetro *b* en bajo ($b \leq 1$), medio ($-1 < b < 1$) y alto ($b \geq 1$); y la cantidad de DIF en baja (≤ 0.40), media (≈ 0.60) y alta (≥ 0.80). Para determinar la fiabilidad de la codificación, dos revisores extrajeron los datos de 5 documentos, los cuales fueron seleccionados de forma aleatoria de entre todos los trabajos. Se calculó la concordancia entre codificadores, alcanzando un valor del índice kappa de 0.99, por lo que la información del resto de documentos fue extraída por un solo codificador.

Obtención de los tamaños del efecto

En el estudio del DIF, se entiende por Error Tipo I a la proporción de veces que un ítem es identificado con DIF cuando en realidad no lo presenta, y por potencia estadística a la proporción de veces que un ítem es identificado correctamente con presencia de DIF. Precisamente estas tasas de detección de cada una de las condiciones de simulación fueron las que se emplearon para el cómputo de los tamaños del efecto. Se procedió transformando las proporciones de detección a *logits*, con sus correspondientes variancias para, a la hora de meta-analizar, determinar el peso asignado a cada una de las condiciones de simulación.

Dada una proporción o tasa de detección *p*, el *logit* de *p* viene dado por:

$$Logit(p) = Ln\left(\frac{p}{1-p}\right)$$

y su correspondiente variancia por:

$$V(Logit) = \frac{1}{np(1-p)}$$

donde *n* es el número de réplicas de la simulación.

Para facilitar la interpretación de los resultados, los datos se convirtieron de nuevo a proporciones, siguiendo la siguiente fórmula:

$$p = \frac{e^x}{1+e^x}$$

donde *x* es el *logit* de *p*.

Análisis de datos

Para obtener el tamaño del efecto medio del conjunto de estudios sobre MH se asignó un peso a cada uno de los tamaños del efecto, el cual viene definido por el inverso de su variancia, bajo un modelo de efectos aleatorios. Se trabajó a un intervalo de confianza del 95% (95% IC).

Para evaluar la homogeneidad de los tamaños del efecto se empleó el estadístico Q (Hedges y Olkin, 1985); la aceptación de la hipótesis nula de homogeneidad indicará que la dispersión de los tamaños del efecto respecto a la media no es mayor que la esperada por el error de muestreo, mientras que su rechazo indicará que la variabilidad entre tamaños del efecto es superior a la esperada por ese error y, por tanto, que pueden estar influyendo otras variables como las características de los estudios, en nuestro caso, las condiciones de simulación.

Para analizar cada uno de los efectos de las variables moderadoras sobre los tamaños del efecto se emplearon dos estrategias bajo el modelo de efectos mixtos mediante el procedimiento de máxima verosimilitud restringida: en el caso de tratarse de variables categóricas se aplicó el método análogo al análisis de la variancia para meta-análisis (Hedges, 1982) y cuando se analizaron variables continuas se empleó la regresión simple ponderada (Hedges y Olkin, 1985).

Los análisis fueron realizados con las macros MeanES, MetaF y MetaReg para SPSS de Wilson (Lipsey y Wilson, 2001) y se trabajó a un nivel de significación de α=0.05.

Resultados

Siguiendo el objetivo principal del presente trabajo, y como se ha comentado anteriormente, se plantearon dos estudios con propósitos distintos: el *estudio 1* se diseñó para explorar la influencia de distintas variables en la tasa de Error Tipo I del MH, y el *estudio 2* para averiguar la influencia de estas variables, incluyendo el Error Tipo I, en la potencia estadística de la prueba. A continuación se exponen los principales resultados obtenidos en ambos estudios, seguidos de un estudio descriptivo de las condiciones donde MH funciona adecuadamente.

Estudio 1. Tasa de Error Tipo I de MH

En este análisis se contemplaron los documentos que cumplieron los criterios de inclusión previamente expuestos, limitándolos a aquellos que se interesaron por la tasa de Error Tipo I. Fueron un total de 17 artículos, los cuales suponen 775 condiciones de simulación que se resumen en la tabla 2.

TABLA 2. DESCRIPTIVOS GENERALES DE LAS CONDICIONES DE SIMULACIÓN

Variable	N	%	Variable	N	%
Modelo de simulación			Parámetro ***b***		
Logístico de 1-p	53	6.8	Bajo	102	21.6
Logístico de 2-p	130	16.8	Medio	257	54.3
Logístico de 3-p	587	75.7	Alto	114	24.1
Unifactorial estricto	5	0.6	Tipo de DIF		
Parámetro ***a***			Uniforme	539	93.1
Bajo	70	12.9	No-uniforme	40	6.9
Medio	276	50.7	Cantidad de DIF		
Alto	198	36.4	Baja	189	73.3
			Media	43	16.7
			Alta	26	10.1
Variable	**N**	**Media (DT)**	**Variable**	**N**	**Media (DT)**
Parámetro ***c***	470	0.14 (0.107)	Número de ítems en test	763	31.60 (12.834)
Sujetos grupo de referencia	759	978.75 (843.036)	Sujetos grupo focal	759	808.20 (778.110)
Razón sujetos R/F	761	1.58 (1.728)	Cantidad de impacto	759	0.64 (0.552)
Porcentaje ítems con DIF	726	7.83 (13.096)			

N=número de datos integrados; DT=desviación típica.

El tamaño del efecto medio, en términos de proporciones, para el Error Tipo I fue de 0.1057 (95% IC: 0.0986-0.1133) y la prueba de homogeneidad entre condiciones de simulación fue estadísticamente significativa ($Q_{(774)}$= 29172.2786; p=.0000). En la tabla 3 se detallan los resultados de los análisis de la variancia para cada una de las variables categóricas. A excepción del valor del parámetro de discriminación, todas las variables resultaron afectar a la tasa de Error Tipo I.

TABLA 3. RESUMEN DE LOS ANÁLISIS DE VARIANCIA PARA LAS VARIABLES CATEGÓRICAS

Variable	N	Q_B(g.l.)	Q_W (g.l.)
Modelo	775	14.4613 (3) *	823.8118 (771)
Parámetro ***a***	544	2.2820 (2)	584.3536 (541)
Parámetro ***b***	473	56.1307 (2) *	517.0608 (470)
Tipo de DIF	579	16.6510 (1) *	614.6213 (577)
Cantidad de DIF	258	18.4749 (2) *	248.8099 (255)

* $p < 0.05$; N: número de datos integrados; Q_B: valor de la prueba de homogeneidad entre-estudios; Q_W: valor de la prueba de homogeneidad intra-estudios

En relación con el efecto de cada una de las variables moderadoras cuantitativas, en la tabla 4 se presentan los resultados hallados en la aplicación de la regresión simple ponderada. En este caso, a excepción del parámetro de pseu-

doazar, todas las variables presentan un efecto significativo sobre la tasa de Error Tipo I; sin embargo, la proporción de variancia explicada por cada una de las variables es relativamente escasa, a excepción del porcentaje de ítems con funcionamiento diferencial (11.35%).

TABLA 4. RESUMEN DE LOS ANÁLISIS DE REGRESIÓN SIMPLE PONDERADA PARA LAS VARIABLES CONTINUAS

Variable	N	C.R.	Q_R (g.l.)	Q_E (g.l.)	R^2
Parámetro ***c***	470	0.7293	1.8705 (1)	495.1085 (468)	0.0038
Número de ítems	763	-0.0081	5.9340 (1) *	813.0279 (761)	0.0072
% ítems con DIF	726	0.0328	99.8323 (1) *	779.5873 (724)	0.1135
N grupo referencia	759	0.0002	17.9654 (1) *	805.1666 (757)	0.0218
N grupo focal	759	0.0002	12.9994 (1) *	805.8427 (757)	0.0159
Razón R/F	761	-0.0610	5.7391 (1) *	810.0977 (759)	0.0070
Impacto	759	0.6276	64.1321 (1) *	808.6823 (757)	0.0735

* $p < 0.05$; N: número de datos integrados; C.R.: coeficiente de regresión; Q_R: valor de la prueba de homogeneidad del modelo de regresión; Q_E: valor de la prueba de homogeneidad del error; R^2: coeficiente de determinación

Estudio 2. Potencia de MH

En este análisis se contemplaron los documentos que cumplieron los criterios de inclusión previamente expuestos, limitándolos a aquellos que se interesaron por la potencia de la técnica. Fueron un total de 11, los cuales suponen 376 condiciones de simulación (ver tabla 5).

TABLA 5. DESCRIPTIVOS GENERALES DE LAS CONDICIONES DE SIMULACIÓN

Variable	N	%	Variable	N	%
Modelo de simulación			Parámetro ***b***		
Logístico de 1-p	38	10.6	Bajo	8	7.0
Logístico de 2-p	104	29.0	Medio	104	91.2
Logístico de 3-p	212	59.1	Alto	2	1.8
Unifactorial estricto	5	1.4	Tipo de DIF		
Parámetro ***a***			Uniforme	222	62.9
Bajo	4	6.0	No-uniforme	131	37.1
Medio	59	88.1	Cantidad de DIF		
Alto	4	6.0	Baja	233	71.7
			Media	59	18.2
			Alta	33	10.2

Variable	N	Media (DT)	Variable	N	Media (DT)
Parámetro ***c***	174	0.082 (0.099)	Número de ítems en test	343	39.20 (13.498)
Sujetos grupo de referencia	343	760.28 (673.560)	Sujetos grupo focal	343	749.78 (738.713)
Razón sujetos R/F	343	1.6723 (2.435)	Cantidad de impacto	339	0.3717 (0.447)
Porcentaje ítems con DIF	343	21.0181 (26.764)			

N=número de datos integrados; DT=desviación típica.

El tamaño del efecto medio para la potencia estadística del MH fue de 0.6554 (95% IC: 0.6300-0.6798) y la prueba de homogeneidad entre condiciones de simulación fue estadísticamente significativa ($Q_{(375)}$= 30183.4974; p=.0000). En la tabla 6 se detallan los resultados de los análisis de la variancia para cada una de las variables categóricas. En este caso, todas las variables moderadoras resultaron afectar a la potencia del MH.

TABLA 6. RESUMEN DE LOS ANÁLISIS DE VARIANCIA PARA LAS VARIABLES CATEGÓRICAS

Variable	N	Q_B (g.l.)	Q_W (g.l.)
Modelo	359	47.0561 (3) *	386.1046 (355)
Parámetro ***a***	67	8.7195 (2) *	65.0034 (64)
Parámetro ***b***	114	6.2500 (2) *	118.3305 (111)
Tipo de DIF	353	31.5708 (1) *	361.6310 (351)
Cantidad de DIF	325	8.8916 (2) *	343.0560 (322)

* $p < 0.05$; N: número de datos integrados; Q_B: valor de la prueba de homogeneidad entre-estudios; Q_W: valor de la prueba de homogeneidad intra-estudios

En relación con el efecto de cada una de las variables moderadoras cuantitativas, en la tabla 7 se presentan los resultados hallados en la aplicación de la regresión simple ponderada. Las variables que resultaron afectar individualmente de forma significativa a la potencia estadística de MH fueron el valor del parámetro *c*, el número de sujetos en el grupo focal, la razón entre am-bos grupos y la tasa de Error Tipo I; sin embargo, en ninguno de los casos, la va-riancia explicada por cada una de las variables fue muy elevada. Sorprendentemente, ni el número de ítems del test ni el porcentaje de ítems con DIF resultaron afectar de forma estadísticamente significativa a la potencia de MH, cuando ambas son variables que influyen en la precisión de la estimación del nivel de habilidad de los sujetos, es decir, en la mayor o menor contaminación del criterio de equiparación de los grupos y, por tanto, en el resultado de la prueba estadística.

TABLA 7. RESUMEN DE LOS ANÁLISIS DE REGRESIÓN SIMPLE PONDERADA PARA LAS VARIABLES CONTINUAS

Variable	N	C.R.	Q_R(g.l.)	Q_E (g.l.)	R^2
Parámetro ***c***	174	-2.7992	7.0439 (1) *	173.5449 (172)	0.0390
Número de ítems	343	-0.0045	0.3949 (1)	361.3853 (341)	0.0011
% ítems con DIF	343	-0.0025	0.4679 (1)	361.2695 (341)	0.0013
N grupo referencia	343	0.0000	0.0006 (1)	361.6994 (341)	0.0000
N grupo focal	343	0.0004	10.9956 (1) *	360.7763 (341)	0.0296
Razón R/F	343	-0.1299	10.9905 (1) *	361.8439 (341)	0.0295
Impacto	339	-0.0509	0.0524 (1)	357.5826 (337)	0.0001
Error Tipo I	219	-2.2765	8.0453 (1) *	230.0195 (217)	0.0338

* $p < 0.05$; N: número de datos integrados; C.R.: coeficiente de regresión; Q_R: valor de la prueba de homogeneidad del modelo de regresión; Q_E: valor de la prueba de homogeneidad del error; R^2: coeficiente de determinación

Con la finalidad de integrar descriptivamente ambos estudios, se asume que una tasa de Error Tipo I, según el criterio de robustez estricto (Bradley, 1978), es adecuada con valores menores de 0.055 y una potencia estadística es apropiada con un valor igual o superior a 0.80 (Cohen, 1988). En la tabla 8 se presentan los descriptivos para cada una de las variables moderadoras en aquellos casos en que se cumplieron estos criterios, teniendo en cuenta los tamaños del efecto observados en cada condición de simulación. Los datos se presentan en términos de porcentajes o de medias y rangos, dependiendo de la naturaleza de las variables.

TABLA 8. DESCRIPTIVOS DE LAS VARIABLES CUANDO SE CUMPLEN CRITERIOS DE EFICACIA

Variables	Error Tipo I	Potencia
Modelo		
Logístico de 1-p	4.0%	16.3%
Logístico de 2-p	18.4%	35.6%
Logístico de 3-p	77.3%	45.2%
Factorial estricto	0.4%	2.9%
Parámetro ***a***		
Bajo	9.4%	0.0%
Medio	52.7%	96.8%
Alto	37.9%	3.2%
Parámetro ***b***		
Bajo	14.7%	11.1%
Medio	60.5%	88.9%
Alto	24.7%	0.0%
Parámetro ***c***	0.13 (0.00-0.25)	0.07 (0.00-0.20)
Número de ítems	31.20 (20-70)	37.86 (20-60)
N grupo referencia	866.21 (75-3000)	802.88 (250-3000)
N grupo focal	710.90 (25-3000)	943.75 (100-4000)
Razón R/F	1.83 (1.00-20.00)	1.31 (0.25-5.00)
Impacto	0.49 (0.0-1.5)	0.35 (0.0-1.0)
Tipo de DIF		
Uniforme	89.5%	56.1%
No-uniforme	10.5%	31.8%
Cantidad de DIF		
Baja	54.5%	66.0%
Media	25.5%	23.7%
Alta	20.0%	10.3%
% ítems con DIF	2.06 (0-15)	15.16 (0-100)
Error Tipo I	-	0.13 (0.025-0.736)

Por tanto, teniendo en cuenta ambos estudios y solamente las variables categóricas, el procedimiento MH como técnica de detección del DIF cumple ambos criterios (Error Tipo I < 0.055 y Potencia ≥ 0.80) en un gran abanico de situaciones; sin embargo, el balance entre criterios se alcanza en un mayor porcentaje trabajando con modelos logísticos de 3-p, valores medios del parámetro de discriminación, valores medios también del parámetro de dificultad, en presencia de DIF uniforme y con una cantidad de DIF baja. En cuanto a las variables continuas, se encontró de nuevo una gran diversidad de escena-

rios donde se cumplieron ambos criterios; no obstante, la media del valor del parámetro *c* se sitúa entre 0.07 y 0.13, el promedio del número de ítems que contiene el test se sitúa aproximadamente entre 30 y 40, el valor medio de sujetos en el grupo de referencia entre 800 y 900, y en el grupo focal entre 700 y 1000, mientras que la razón media entre los tamaños muestrales de ambos grupos es superior a 1, lo que indica que MH funciona mejor cuando el grupo de referencia contiene más sujetos que el grupo focal. Además, el valor medio de impacto se sitúa en torno a una diferencia entre medias de ambos grupos de 0.50 unidades de desviación típica y, por último, el porcentaje medio de ítems con DIF en el test entre 2% y 15%, aproximadamente.

Discusión

En este trabajo se ha presentado un estudio meta-analítico de los artículos que han explorado el Error Tipo I y/o la potencia estadística del procedimiento MH en la detección del DIF, especificando la influencia de distintas variables moderadoras sobre estas tasas de detección. Además se han integrado, de forma descriptiva, ambos estudios con la finalidad de determinar las condiciones bajo las cuales se controlan la tasa de error Tipo I y la potencia de MH. En consecuencia, la información aquí incluida puede ser de gran utilidad tanto para los investigadores que se interesan por el funcionamiento de la técnica MH como para los que desean aplicar este procedimiento como técnica de detección del DIF en contextos empíricos de medida. Sin embargo, aún quedan cuestiones sobre el MH por explorar.

Por un lado, tal y como se ha comentado anteriormente, para evitar la dependencia entre tamaños del efecto, en este trabajo solamente se han tenido en cuenta aquellos estudios que trabajaron a un nivel de significación estadística del 0.05 y que no utilizaron procedimientos de purificación. En este sentido, en investigaciones futuras sería interesante explorar de nuevo la tasa de Error Tipo I y la potencia de aquellos trabajos que, por un lado, emplean otros niveles de significación y, por otro, aplican procedimientos de purificación bietápica o iterativa. En este último caso, algunos estudios realizados hasta el momento parecen indicar que la aplicación de procedimientos de purificación supone una cierta mejoría en las tasas de detección para algunas condiciones, mientras que para otras no (Clauser, Mazor y Hambleton, 1993; Fidalgo, Mellenbergh y Muñiz, 1998; Miller y Oshima, 1992; Wang y Su, 2004a); por tanto, habrá que explorar el efecto de estas variables de nuevo mediante estudios meta-analíticos.

Por otro lado, sería muy interesante explorar, con meta-análisis, el funcionamiento de MH como procedimiento de detección del DIF cuando se trabaja con ítems de respuesta politómica. La propuesta de extensión de MH para la detección del funcionamiento diferencial en este tipo de ítems viene de manos de Mantel (1963); aunque se han publicado más estudios simulados en el caso de ítems dicotómicos, existen un gran número de trabajos dedicados al otro formato de respuesta (Chang, Mazzeo y Roussos, 1996; Kristjansson,

Aylesworth, McDowell y Zumbo, 2005; Su y Wang, 2005; Wang y Su, 2004b; Zwick, Donoghue y Grima, 1993; Zwick y Thayer, 1996; Zwick, Thayer y Mazzeo, 1997), por lo que en futuros estudios se investigará en profundidad este aspecto.

REFERENCIAS

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

Chang, H.H., Mazzeo, J. & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33(3),* 333-353.

Clauser, B., Mazor, K. & Hambleton, R.K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6(4),* 269-279.

Cohen J. (1988). *Statistical power analysis* (2nd Edition). Hillsdale, NJ: LEA.

Fidalgo, A.M., Mellenbergh, G.J. & 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),* 209-218.

Gómez, J. & Hidalgo, M.M. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: Una revisión metodológica. *Anuario de Psicología, 74,* 3-32.

Hedges, L.V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490-499.

Hedges, L.V. & Olkin I. (1985). Statistical methods for meta-analysis. Orlando: Academic Press.

Holland, P.W. (1985). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology, 18,* 449-493.

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

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

Lipsey, M.W. & Wilson, D.B. (2001). *Practical meta-analysis.* Applied Social Research Methods Series, volume 49. California, USA: Sage Publications, Inc.

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

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

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

Miller, M.D. & Oshima, T.C. (1992). Effect of simple size, number of biased ítems, and magnitude of bias on two-stage item bias estimation method. *Applied Psychological Measurement, 16(4),* 381-388.

Narayanan, P. & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20(3),* 257-274.

Su, Y.H. & Wang, W.C. (2005). Efficiency of the Mantel, Generalized Mantel-Haenszel, and Logistic Discriminant Function Analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education, 18(4),* 313-350.

Wang, W.C. & Su, Y.H. (2004a). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel Method. *Applied Measurement in Education, 17(2),* 113-144.

Wang, W.C. & Su, Y.H. (2004b). Factors influencing the Mantel and the Generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28(6),* 450-480.

Zwick, R., Donoghue, J.R. & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30(3),* 233-251.

Zwick, R. & Thayer, D.T. (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. & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10(4),* 321-344.