Psicothema
2000. Vol. 12, Supl. nº 2, pp. 220-225

ISSN 0214 - 9915 CODEN PSOTEG

Detección del funcionamiento diferencial de los ítems no uniforme: comparación de los métodos Mantel-Haenszel y regresión logística

Doris Ferreres Traver, Ángel M. Fidalgo Aliste* y José Muñiz*
Universidad de Valencia y * Universidad de Oviedo

El presente trabajo de simulación pretende identificar bajo qué condiciones el estadístico Mantel-Haenszel en su forma estándar, la modificación propuesta por Mazor, Clauser y Hambleton (1994), y la técnica de la Regresion Logística (RL) son capaces de detectar el funcionamiento diferencial de los ítems no uniforme. Los datos simulados en este estudio procedieron de datos empíricos (Ferreres, 1998). Los parámetros utilizados en la generación de los datos fueron tomados de una aplicación de la escala Habilidad Mental Verbal (HMV), parte integrante de la batería BADYG Medio (Yuste, 1988). Las variables manipuladas fueron: el tamaño muestral, la presencia o no de impacto, el tipo de FDI no uniforme, el porcentaje de ítems con FDI en el test y el nivel de significación. La eficacia de los diferentes métodos se analizó en términos de la potencia de prueba y la tasa de error de Tipo I que presentaron a través de 100 replicaciones. Los resultados contribuyen a clarificar el comportamiento de dichos métodos en presencia del FDI no uniforme. Asimismo, se presentan conclusiones a considerar para futuras investigaciones en la detección del FDI no uniforme.

Detection of non-uniform DIF: Mantel-Haenszel and logistic regression methods. The present simulation study compared the performance under different conditions of the Mantel-Haenszel (MH) statistic, the variation of the MH statistic proposed by Mazor, Clauser and Hambleton (1994), and the logistic regression (LR) on detection of nonuniform differential item functioning (DIF). Simulated data were obtained from a real study (Ferreres,1998). Item parameters used in the simulation are from a aplication of Verbal Mental Ability Scale of Middle BADYG (Yuste,1988). Variables manipulated were: sample size, differences in the group trait level averages, type of nonuniform DIF, proportions of DIF items and signification level. These methods were evaluated by examining changes in the power and the Type I error rate over 100 replications. Results obtained have contributed to clarify the utility of these statistical methods in the presence of nonuniform DIF. Conclusions to consider in future studies of the deteccion of nonuniform DIF are given.

En los últimos tiempos, han ido apareciendo una gran variedad de métodos y técnicas estadísticas que han posibilitado la detección y evaluación del funcionamiento diferencial de los ítems (FDI). Para una revisión metodológica puede consultarse los textos de Fidalgo (1996), Gómez e Hidalgo (1997), y Millsap y Everson (1993). La mayor parte de los artículos y trabajos relacionados con este área de investigación se han ocupado preferentemente de averiguar qué métodos estadísticos y bajo qué condiciones detectan adecuadamente la presencia del funcionamiento diferencial de los ítems. Además, se han utilizado casi siempre estudios de simulación (Mazor, Clauser y Hambleton, 1992, 1994; Miller y Oshima, 1992; Swaminathan y Rogers, 1990; Rogers y Swaminathan, 1993; Uttaro y Millsap, 1994), centrados en la detección y evaluación del FDI uniforme, quedando en un segundo plano todo lo relacionado con el FDI no uniforme. Esto ha conducido a un mayor desarrollo de los métodos de análisis del FDI uniforme. Esta situación es un claro reflejo de la mayor presencia del FDI uniforme en los tests estandarizados, aunque el FDI no uniforme también ha sido localizado en ciertas ocasiones con datos reales (Hambleton y Rogers, 1989; Linn, Levine, Hastings y Wardrop, 1981; Mellenbergh, 1982). Pese a ello, dentro de esta orientación se han ofrecido algunas aportaciones interesantes acerca de la detección y evaluación del FDI no uniforme. Destacan entre otras, una modificación de cálculo del estadístico Mantel-Haenszel (Mazor, Clauser y Hambleton, 1994) y la técnica de la Regresión Logística (RL) (Swaminathan y Rogers, 1990).

Antes de introducir las citadas técnicas estadísticas, conviene destacar la clasificación del FDI no uniforme propuesta por Swaminathan y Rogers (1990). En su artículo, estos autores distinguen entre FDI no uniforme simétrico y mixto. El FDI no uniforme simétrico ocurre en aquellos casos en que las diferencias en la probabilidad de contestar correctamente el ítem entre los dos grupos se cancelan, esto es, cuando el parámetro *b* se mantiene constante y el parámetro *a* varía entre ambos grupos ($a_R \neq a_F$ y $b_R = b_F$), y el FDI no uniforme mixto cuando las mencionadas diferencias no se anulan, es decir, cuando los parámetros *a* y *b* son desiguales para

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

ambos grupos ($a_R \neq a_F$ y $b_R \neq b_F$). Bajo la óptica de la TRI, el FDI no uniforme simétrico quedaría representado por un cruce central de la curva característica del ítem (CCI) del grupo de referencia y del grupo focal dentro del rango establecido, y el FDI no uniforme mixto, por un cruce no simétrico de la CCI del grupo de referencia y del grupo focal. Posteriormente, Li y Stout (1993) acuñaron los términos «no direccional» (FDI no uniforme simétrico) y «unidireccional» (FDI no uniforme mixto) para los dos tipos de FDI no uniforme.

En la actualidad, el procedimiento MH (Holland y Thayer, 1986) es el método por excelencia en la detección del funcionamiento diferencial. Su uso habitual deriva de la simplicidad de su cálculo e interpretación, y sus buenos resultados con tamaños muestrales pequeños, además de ofrecer una cuantificación de la magnitud del FDI (cociente de razones común, α_{MH}), y un test de significación estadística (estadístico ji-cuadrado MH, χ^2_{MH}). Sin embargo, una de las mayores críticas a este método ha sido su incapacidad para detectar el FDI no uniforme (Swaminathan y Rogers, 1990; Rogers y Swaminathan, 1993). Una forma de erradicar esta deficiencia ha sido la modificación en el cálculo del estadístico MH propuesta por Mazor, Clauser y Hambleton (1994). Esta nueva variante consiste en calcular el estadístico χ^2_{MH} de forma separada en el grupo de sujetos con las puntuaciones más bajas (en concreto, los sujetos con una puntuación total menor o igual a la media de la distribución de las puntuaciones en el test) y en el grupo de sujetos con las mayores puntuaciones (sujetos con una puntuación total mayor que la media de la distribución de las puntuaciones en el test). Algunos estudios de simulación han aplicado esta nueva variación del MH para detectar el FDI no uniforme, y todo indica que es una alternativa viable y eficaz en la detección de este tipo de FDI, aunque a costa de incrementar la tasa de error de Tipo I (Fidalgo, 1996; Fidalgo y Mellenbergh, 1995).

Por su parte, Swaminathan y Rogers (1990) propusieron la regresión logística como método de análisis en la detección del FDI uniforme y no uniforme en ítems dicotómicos. En particular, este modelo permite predecir la probabilidad de una respuesta correcta a un determinado ítem en función del nivel de habilidad del examinado (θ: puntuación total en el test), su grupo de pertenencia (G), y del término θ x G, que representa la interacción entre el nivel de habilidad del examinado y su grupo de pertenencia, siendo este último término el que, en caso de resultar significativo, indicaría la presencia del FDI no uniforme. La ventaja de esta técnica frente al estadístico MH radica en su mayor potencia a la hora de identificar adecuadamente el FDI no uniforme, aunque su análisis es más complejo y costoso de realizar (Rogers y Swaminathan, 1993).

En términos generales, los estudios de simulación realizados coinciden en señalar que: a) el MH estándar es eficaz en la detección del FDI NU mixto, pero incapaz de identificar el FDI NU simétrico; b) la nueva variante del método MH se presenta como una alternativa eficaz para detectar el FDI NU simétrico y mixto, aunque a costa de incrementar la tasa de error de Tipo I; c) la RL ofrece una potencia adecuada para la detección del FDI NU simétrico y mixto, debido a la inclusión del término θ x G en su modelo, pero por contra presenta tasas de error de Tipo I fuera de los límites nominales establecidos, y por último, d) los resultados conseguidos denotan un aumento de las clasificaciones correctas del FDI cuando se utilizan procedimientos iterativos frente a la evaluación del FDI en un solo análisis (Fidalgo, 1996; Fidalgo y Mellenbergh, 1995; Gómez y Navas, 1996; Lautenschlager, Flaherty y Park, 1994).

Teniendo presente lo dicho con anterioridad, hemos de destacar que no existen trabajos que determinen el grado de acuerdo existente en la detección del FDI no uniforme entre la modificación de cálculo propuesta del estadístico χ^2_{MH} y la regresión logística. Este será uno de los objetivos del presente estudio. Más concretamente, este trabajo intenta averiguar bajo diversas condiciones la capacidad del estadístico ji-cuadrado Mantel-Haenszel (MH) en su forma estándar, de la modificación propuesta por Mazor, Clauser y Hambleton (1994), y de la técnica de Regresion Logística (RL) para detectar adecuadamente la presencia del FDI no uniforme simétrico y mixto, y esto utilizando como parámetros en las simulaciones estimaciones obtenidas en muestras españolas (Ferreres, 1998).

Método

Generación de los datos

Con la intención de que las condiciones creadas fueran lo más representativas posibles de una situación real, los parámetros utilizados en la generación de los datos fueron tomados de los datos obtenidos por Ferreres (1998) en la aplicación de la escala Habilidad Mental Verbal (HMV), parte integrante de la batería BADYG Medio (Yuste, 1988). Los valores de los 40 ítems que componen dicha escala en el grupo de referencia, y que se utilizaron como parámetros en la generación, pueden observarse en la Tabla 1.

En la generación de los datos se utilizó un modelo logístico de tres parámetros y las siguientes condiciones para la distribución de la habilidad en los grupos: el grupo de referencia poseía una habilidad con distribución normal N (0,1), mientras que se consideraron dos grupos focales: el primero con la misma distribución de habilidad que el grupo de referencia, y el segundo con una desviación típica por debajo de la media del grupo de referencia N (-1,1). De esta manera, se generaron las respuestas de los examinados a un test compuesto de 46 ítems (40 ítems originales más 6 ítems sesgados). Los 6 ítems sesgados (ítems 41 al 46) objeto de estudio fueron añadidos a los 40 ítems que formaban el test. El parámetro c fue idéntico para ambos grupos en todos los ítems, siendo su valor de 0.2. Igualmente, el valor de los parámetros de los ítems sin FDI fue el mismo para el grupo de referencia y focal. Los parámetros de los 6 ítems sesgados con FDI no uniforme se obtuvieron tras cruzar dos niveles de *a* (0.5 y 0.75) con tres niveles de *b* (-1.5, 0 y 1.5). El funcionamiento diferencial se generó variando, bien el parámetro *a* (FDI no uniforme simétrico), bien los parámetros *a* y *b* (FDI no uniforme mixto) en el grupo focal, de forma que la magnitud del FDI calculada, aplicando la medida de área exacta sin signo (Raju, 1988, 1990), fuese de 0.5. La Tabla 2 muestra los valores de los parámetros de los 6 ítems sesgados para los grupos de referencia y focal.

Por otra parte, de los 40 ítems del test se seleccionaron aleatoriamente 6 (ítems 10, 21, 22, 28, 35, 36) con el fin de poder conseguir un test con el 26% de ítems con FDI no uniforme. De esta forma, cuando el porcentaje de ítems con FDI en el test era del 13%, sólo los 6 ítems añadidos al test presentaban FDI no uniforme simétrico o mixto, y cuando el porcentaje de ítems con FDI en el test era del 26%, los 6 ítems añadidos al test más los 6 ítems seleccionados del test, en total 12 ítems, presentaban FDI no uniforme simétrico o mixto.

Variables manipuladas

En este estudio se manipuló el tamaño muestral, la presencia o no de impacto, el tipo de FDI no uniforme, el porcentaje de ítems con FDI en el test y el nivel de significación. En el tamaño muestral se manejaron dos condiciones: a) 650 (grupo de referencia) y 500 (grupo focal), y b) 350 (grupo de referencia) y 200 (grupo focal). En la distribución de la habilidad se manipularon otros dos niveles: a) Igual distribución de la habilidad en los grupos de referencia y focal (No impacto) y b) Diferente distribución de la habilidad en ambos grupos (Impacto). El tipo de FDI no uniforme fue simétrico y mixto. El porcentaje de FDI en el test fue del 13% y 26%. El número total de condiciones fue de 16 (2 tamaños de muestra x 2 niveles de habilidad x 2 tipos de FDI x 2 porcentajes de ítems con FDI), de cada una de las cuales se hicieron 100 replicaciones. Los datos así obtenidos se analizaron mediante el procedimiento MH estándar, su modificación y la RL, bajo 2 niveles de significación el 0.05 y el 0.01.

Tabla 1
Parámetros de los ítems en el grupo de referencia

Item nº	*b*	*a*	Item nº	*b*	*a*
1	-1.396	0.846	21	1.257	0.606
2	-0.759	0.678	22	0.358	0.550
3	-0.921	0.824	23	-0.050	0.985
4	-0.756	0.788	24	1.520	0.508
5	-0.975	0.587	25	0.897	0.360
6	-0.793	0.767	26	1.578	0.592
7	-0.987	0.467	27	0.241	0.509
8	0.024	0.556	28	0.017	0.789
9	-0.360	0.753	29	0.097	0.517
10	-0.190	0.857	30	0.487	0.782
11	-1.376	1.026	31	0.825	1.020
12	-1.196	0.816	32	1.757	0.536
13	-1.135	0.787	33	0.737	0.695
14	-0.732	0.812	34	0.964	0.591
15	-0.637	0.725	35	2.600	0.542
16	-0.285	0.735	36	1.036	0.531
17	0.632	0.640	37	0.763	0.492
18	-0.569	0.647	38	0.990	0.529
19	0.076	0.717	39	1.783	0.392
20	0.736	0.727	40	2.968	0.311

Tabla 2
Parámetros de los ítems sesgados en el grupo de referencia y local

			FDI NO UNIFORME		
	GRUPO REFERENCIA		Simétrico	Mixto GRUPO FOCAL	
Item	*b*	*a*	*a*	*b*	*a*
10	-0.190	0.857	2.4973	0.44	0.900
21	1.257	0.606	1.1315	0.85	1.000
22	0.358	0.550	0.9508	-0.27	0.587
28	0.017	0.789	1.9960	-0.56	1.200
35	2.600	0.542	0.9271	1.975	0.550
36	1.036	0.531	0.8954	1.55	0.750
41	-1.5	0.5	0.8106	-1.00	0.700
42	-1.5	0.75	1.7639	-1.00	1.300
43	0	0.5	0.8106	0.50	0.700
44	0	0.75	1.7639	0.50	1.300
45	1.5	0.5	0.8106	1.00	0.700
46	1.5	0.75	1.7639	1.00	1.300

Detección del FDI

Los programas utilizados fueron el EZDIF (Waller, 1998) y el MHDIF (Fidalgo, 1994). El primero implementa el MH estándar y la RL para la detección del FDI uniforme y no uniforme con datos dicotómicos. Por su parte, el programa MHDIF permite computar el MH estándar y la modificación propuesta por Mazor, Clauser y Hambleton (1994) para el cálculo del FDI no uniforme. Para evaluar el FDI no uniforme en la regresión logística se calculó el término θ x G que representa la interacción entre el nivel de habilidad del sujeto y su grupo de pertenencia. Señalar que tanto el programa MHDIF como el EZDIF implementan procedimientos bietápicos de purificación de las puntuaciones totales en el test, esto significa que ambos procedimientos, en una segunda etapa vuelven a recalcular los estadísticos para detectar el FDI, utilizando para el cálculo de la puntuación total en el test los ítems que no fueron identificados como sesgados en el primer análisis.

Además, se aplicaron dos formas diferentes de purificar las puntuaciones totales en el test con la nueva propuesta del MH. El primer procedimiento (MHNU1) procedió a eliminar los ítems que presentaban FDI en el grupo de menor habilidad o en el de mayor habilidad en el primer análisis, posteriormente la muestra fue de nuevo dividida en el grupo de menor habilidad y el grupo de mayor habilidad de acuerdo a la media de las puntuaciones purificadas, y los estadísticos MH fueron nuevamente recalculados. El segundo procedimiento (MHNU2) eliminó los ítems que presentan FDI en el primer análisis en el grupo de menor habilidad y posteriormente recalculó los estadísticos MH en dicho grupo. Del mis-

Tabla 3
Porcentaje medio de identificaciones correctas de los ítems con FDI analizados (ítems 41 a 46), en cada una de las condiciones simuladas y para cada método de análisis con niveles de significación de 0.05 y 0.01

α= 0.05										
FDI NU	Impacto	% FDI	NGR=650 NGF=500				NGR=350 NGF=200			
			MH	MHNU1	MHNU2	RL	MH	MHNU1	MHNU2	RL
S	NO	13	18	76	77	73	10	61	64	65
		26	17	77	78	74	14	65	65	66
	SI	13	58	70	71	91	43	59	57	74
		26	55	69	69	90	43	57	58	75
M	NO	13	62	66	67	62	44	58	59	54
		26	60	67	69	65	45	60	59	58
	SI	13	73	74	73	80	59	63	62	65
		26	72	73	72	80	58	61	61	65
α=0.01										
FDI NU	Impacto	% FDI	NGR=650 NGF=500				NGR=350 NGF=200			
			MH	MHNU1	MHNU2	RL	MH	MHNU1	MHNU2	RL
S	NO	13	13	66	66	68	5	50	50	57
		26	13	65	65	68	6	54	53	59
	SI	13	48	60	58	82	33	50	49	58
		26	44	59	58	78	34	48	49	64
M	NO	13	51	59	59	55	34	50	51	50
		26	53	59	60	57	33	49	50	51
	SI	13	64	63	63	70	51	55	55	50
		26	63	63	62	71	49	52	52	51

mo modo, se procedió a realizar este procedimiento para el grupo de mayor habilidad.

Resultados

La eficacia de los diferentes métodos empleados se analizó atendiendo su potencia de prueba, es decir, a la proporción (que transformaremos a porcentajes) de ítems con FDI correctamente identificados (IC), así como a la tasa de ítems que sin presentar FDI fueron detectados como tales, es decir, la proporción de falsos positivos (FP) identificados a través de las 100 replicaciones realizadas. Los resultados obtenidos aparecen resumidos teniendo en cuenta los dos niveles de significación considerados: 0.05 y 0.01.

Identificaciones correctas

En la Tabla 3 podemos apreciar los porcentajes medios de las identificaciones correctas detectados en cada condición manipulada a lo largo de las 100 replicaciones para el método MH estándar, las dos formas de purificación aplicadas a la variación del MH propuesta por Mazor y cols. (1994) (MHNU1 y MHNU2), y la regresión logística (RL).

Falsos positivos

En la Tabla 4, podemos observar las tasas de error de Tipo I obtenidas en cada una de las condiciones manipuladas. A simple vista, se podía distinguir que la tasa de falsos positivos (FP) variaba en función del impacto y el nivel de significación seleccionado. De forma que el mayor porcentaje de FP correspondía a la RL, siendo éstos sustancialmente elevados cuando el impacto estaba presente y el nivel de significación considerado fue del 0.05. Por contra, las tasas de error de Tipo I más bajas fueron siempre para el procedimiento MH estándar, aunque en algunos niveles fueron ligeramente superiores (0.06) sobre la tasa nominal correspondiente ($\alpha = 0.05$). Por su parte, la nueva variación del MH también presentó ciertos incrementos en la tasa de falsos positivos, siendo éstos más sustanciales en el nivel de significación del 0.05.

Tipo de ítem

En general, la tasa de IC más elevada para todos los métodos correspondió a los ítems con dificultad media (b=0), y a los ítems con baja dificultad y alta discriminación. Por contra, los ítems peor identificados por todos los métodos fueron los ítems más difíciles, especialmente para el MH estándar. En relación con la variable discriminación, los resultados mostraron que cuanto mayor es el índice de discriminación del ítem más fácil es identificarlo, a excepción de los ítems más difíciles. Al respecto, conviene señalar que la RL se mostró eficaz en la detección del FDI NU simétrico en ítems con elevada dificultad, en especial cuando los ítems poseían niveles bajos de discriminación (en torno al 0.70 ó 0.87%).

Discusión

El objetivo principal de este trabajo ha sido averiguar la utilidad de distintos métodos estadísticos (el MH estándar, la modificación propuesta por Mazor, Clauser y Hambleton (1994), y la Regresión Logística) en la detección del funcionamiento diferencial NU bajo diversas condiciones, utilizando datos procedentes del ámbito aplicado (Ferreres, 1998) como parámetros en la simulación.

Uno de los primeros resultados encontrados que concuerda con los de otros estudios, es que el MH estándar posee escasa potencia de prueba en la detección del FDI NU simétrico (Fidalgo,1996; Rogers y Swaminathan,1993; Uttaro y Millsap,1994; Narayanan y Swaminathan,1996). En concreto, este método presenta un porcentaje de IC del 32,3% al nivel de significación del 0.05; y del 24,6% al nivel del 0.01. Sin embargo, cuando el FDI NU es mixto ofrece una potencia de prueba aceptable. Precisamente, la tasa de IC obtenida en este estudio es comparable a la ofrecida por Mazor, Clauser y Hambleton (1994), siendo como promedio del 67% y 58% con un alfa del 0.05 y 0.01, respectivamente, para la condición de mayor tamaño muestral (N_{GR}=650/N_{GF}=500). Respecto a la nueva variante propuesta, destacar que ambos métodos (MH-

Tabla 4
Promedio de la tasa de error de Tipo I en los ítems con FDI analizados (ítems 41 a 46), en cada una de las condiciones simuladas y para cada método de análisis con niveles de significación de 0.05 y 0.01

			α= 0.05							
FDI NU	**NGR/NGF**	**Impacto**	**13 % FDI**				**26 % FDI**			
			MH	**MHNU1**	**MHNU2**	**RL**	**MH**	**MHNU1**	**MHNU2**	**RL**
S	650/500	NO	.05	.07	.08	.10	.06	.08	.08	.14
		SI	.05	.09	.08	.13	.05	.08	.09	.13
	350/200	NO	.04	.07	.07	.07	.05	.07	.07	.09
		SI	.04	.07	.07	.10	.04	.07	.08	.11
M	650/500	NO	.04	.07	.08	.05	.04	.08	.08	.06
		SI	.05	.09	.08	.12	.06	.10	.08	.12
	350/200	NO	.04	.06	.07	.05	.03	.06	.07	.05
		SI	.05	.07	.07	.10	.04	.08	.08	.10
			α=0.01							
FDI NU	**NGR/NGF**	**Impacto**	**13 % FDI**				**26 % FDI**			
			MH	**MHNU1**	**MHNU2**	**RL**	**MH**	**MHNU1**	**MHNU2**	**RL**
S	650/500	NO	.01	.01	.01	.03	.01	.02	.02	.04
		SI	.01	.02	.02	.05	.01	.02	.02	.05
	350/200	NO	.01	.01	.01	.02	.01	.01	.02	.02
		SI	.01	.01	.01	.03	.01	.01	.01	.03
M	650/500	NO	.01	.02	.01	.01	.01	.02	.02	.01
		SI	.01	.02	.02	.04	.01	.02	.02	.04
	350/200	NO	.01	.01	.01	.01	.01	.01	.01	.01
		SI	.01	.01	.01	.03	.01	.02	.02	.03

Tabla 5
Porcentaje medio de identificaciones correctas de los ítems con FDI analizados (ítems 41 a 46) a lo largo de todas las condiciones en función de los parámetros *b* y *a* de los mismos

	FDI NU Simétrico								FDI NU Mixto							
	MH		MHNU1		MHNU2		RL		MH		MHNU1		MHNU2		RL	
	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01
I41:↓ *b* ↓ *a*	47	32	46	24	51	24	43	21	28	12	35	17	33	18	15	04
I42:↓ *b* ↑ *a*	52	50	100	100	100	100	100	100	100	99	100	100	100	100	100	100
I43:= *b* ↓ *a*	36	24	96	89	96	88	99	96	83	70	98	92	98	94	94	87
I44:= *b* ↑ *a*	46	38	100	100	100	100	100	99	94	88	100	100	100	100	98	92
I45:↑ *b* ↓ *a*	08	02	50	25	49	23	87	70	47	28	50	27	50	26	51	32
I46:↑ *b* ↑ *a*	05	02	09	02	08	02	27	14	06	01	10	02	10	03	38	26

NU1 y MHNU2) presentan una potencia razonable para detectar el FDI NU simétrico (tasa global de IC= 61,75%) y mixto (tasa global de IC= 60,8%). Por contra, presenta como desventaja un elevado número de falsos positivos, resultado acorde con el obtenido en otros estudios (Fidalgo, 1996; Fidalgo y Mellenbergh, 1995). Asimismo, la RL presenta una elevada potencia de prueba en la detección del FDI NU: El porcentaje de IC para el FDI NU simétrico fue del 71,5 %; y del 61,5% para el FDI NU mixto. Sin embargo, esta elevada tasa de IC estuvo acompañada por un aumento muy elevado en el número de FP (véase la Tabla 4), siendo su proporción media del 0.10 y del 0.03 para los dos niveles de significación de 0.05 y 0.01. Este mismo resultado es confirmado por otros estudios (Narayanan y Swaminathan,1996; Swaminathan y Rogers,1990; Hidalgo y López, 1997).

Otro de los hallazgos de este estudio ha sido el escaso efecto que el porcentaje de ítems con FDI en el test tiene en las tasas de IC y FP calculadas para ambos tipos de FDI NU. Este resultado viene respaldado por los estudios de Hidalgo y López (1997), Swaminathan y Rogers (1990), y parcialmente por Narayanan y Swaminathan (1996). La explicación de este resultado, como algunos de los autores citados anteriormente han hecho notar, podría deberse a la utilización de procedimientos bietápicos. En cuanto al tamaño muestral, se encontró, como cabría esperar, que el porcentaje de IC y FP se incrementó al aumentar el número de examinados en los grupos de referencia y focal. Esta circunstancia se mantuvo para todos los procedimientos y condiciones manipuladas en el estudio, y es ampliamente corroborado por la mayoría de los estudios realizados (Hidalgo y López, 1997; Narayanan y Swaminathan, 1996; Rogers y Swaminathan, 1993; Swaminathan y Rogers, 1990).

Mas controvertido es el efecto que el impacto tiene sobre la potencia de prueba y la tasa de error de Tipo I. La presencia de impacto, esto es, de diferentes distribuciones de la habilidad entre los grupos comparados, aumenta la tasa de IC para el MH estándar y la RL cuando el FDI NU es simétrico, y para todos los métodos de análisis cuando el FDI NU es mixto. Este resultado es aplicable a ambos niveles de significación. También, la tasa de error de Tipo I estuvo afectada por el impacto, provocando un aumento generalizado de FP en todos los métodos, siendo especialmente acentuado para la RL y los métodos MHNU2 y MHNU1 al nivel de significación del 0.05. Estos resultados están en consonancia con lo aportado por Fidalgo (1996), a excepción de los resultados conseguidos para el MH estándar, que en su estudio presentó menores tasas de IC con el FDI NU simétrico y mayor porcentaje de FP con ambos tipos de FDI NU cuando el impacto estuvo presente. En cambio, Rogers y Swaminathan (1993) no lograron encontrar diferencias significativas en función del impacto para el MH estándar. Este resultado es respaldado por Hidalgo y López (1997), pero en este caso para la RL. Por su parte, Narayanan y Swaminathan (1996) en su estudio declararon que el impacto disminuía la tasa de IC pero aumentaba la de FP, siendo las diferencias entre impacto y no impacto para la tasa de IC del 2% en el MH estándar y del 14% en la RL.

El siguiente punto a destacar es el efecto diferencial que la magnitud de los parámetros *a* y *b* ejercen sobre la tasa de IC y FP calculados en ambos tipos de FDI. En general, las tasas más elevadas de IC para todos los métodos corresponden a los ítems de dificultad media (b=0), y a los ítems de baja dificultad y alta discriminación, aunque las tasas de identificación más elevadas para el MH estándar cuando el FDI NU es simétrico son únicamente del 40% o el 50%. Este último resultado discrepa con lo encontrado por Fidalgo (1996) y Rogers y Swaminathan (1993), donde los ítems peor identificados por el MH estándar cuando el FDI NU es simétrico son los ítems de dificultad media. Por contra, los ítems peor identificados en el estudio son los ítems con mayor dificultad, especialmente para el MH estándar. No obstante, para este tipo de ítems la RL es capaz de ofrecer elevadas tasas de identificación (entre el 0.70 y 0.87%), especialmente cuando el FDI NU es simétrico y los ítems poseen bajos niveles de discriminación. Respecto a la discriminación de los ítems destacar que, para ambos tipos de FDI NU, cuanto mayor es el índice de discriminación del ítem más fácilmente es identificado, a excepción de los ítems más difíciles. En general, estos resultados concuerdan con lo obtenido por Fidalgo (1996), y Rogers y Swaminathan (1993).

En resumen, los resultados aportados en este estudio llevan a preferir para la detección del DIF no uniforme la variación del procedimiento MH propuesta por Mazor, Clauser y Hambleton (1994), en cualquiera de las dos formas de purificación implementadas (MHNU1 y MHNU2), sobre el resto de los procedimientos utilizados. Frente al MH estándar, por la poca potencia estadística de este último en la detección del FDI NU simétrico, y frente a la RL, por el elevado número de FP de la misma. Esta circunstancia, junto con los dos niveles de significación utilizados en el estudio, nos permite sugerir la utilización de niveles de significación bajos (del 0.01), cuando el coste de elaboración de los ítems sea elevado y se quiera disminuir la probabilidad de eliminar una buena cantidad de ítems incorrectamente. Por otra parte, estos mismos métodos ofrecen un comportamiento óptimo en la detección del FDI no uniforme de los ítems con niveles medios de dificultad, lo que reafirma su utilidad como método de detección del FDI cuando el test a analizar es un test de aptitud dado que su construcción técnicamente se basa en la inclusión de ítems en su mayoría con alta discriminación y dificultad media.

En futuras investigaciones sería aconsejable la comparación de los procedimientos aquí presentados con otros métodos de evaluación del FDI NU, como por ejemplo, el Crossing SIBTEST (Li y Stout, 1993), ó el análisis de residuales (Prieto y Barbero, 1996). Además sería aconsejable la manipulación de aquellas variables con resultados contradictorios —como es el caso del impacto—, y la introducción de variables poco estudiadas hasta el momento pero de suma importancia a la hora de proporcionar información acerca de cómo detectar más adecuadamente la presencia del funcionamiento diferencial no uniforme de los ítems, como por ejemplo las características de los ítems o la longitud del test.

Referencias

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

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

Fidalgo, A.M. y Mellenbergh, G.J. (1995*). Evaluación del procedi - miento 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. (1994). MHDIF: A computer program for detecting uniform and nonuniform differential item functioning with the Mantel-Haenszel procedure*. Applied Psychological Measurement, 18*, 3, 300.

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

Gómez, J. y Navas, M.J. (1996). Detección del funcionamiento diferencial de los ítems mediante regresión logística: purificación paso a paso de la habilidad.*Psicológica, 17*, 2, 397-411.

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

Hidalgo, M.D. y López Pina, J.A. (1997). Comparación entre las medidas de área, el estadístico de Lord y el análisis de regresión logística en la evaluación del funcionamiento diferencial de los ítems. *Psicothema,* 9, 417-431.

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

Lautenschlager, G.J.; Flaherty, V.L. y Park, D.G. (1994). IRT differential item functioning: An examination of ability scale purifications. *Edu - cational and Psychological Measurement, 54,* 1, 21-31. Li, H. y Stout, W.F. (1993). *A new procedure for detection of crossing DIF/bias.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Linn, R.L.; Levine, M.V.; Hastings, C.N. y Wardrop, J.L. (1981). Item bias in a Test of Reading Comprehension. *Applied Psychological Measu - rement,* 5, 159-173.

Mazor, K.; Clauser, B. y Hambleton, R.K. (1992). The effect on sample size on the functioning of the Mantel-Haenszel statistic*. Educational and Psychological Measurement,* 52, 443-451.

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

Mellenbergh, G.J. (1982). Contingency table models for assessing item bias. *Journal of Educational Measurement,* 7, 105-108.

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*, 4, 381-388.

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

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

Prieto, P. y Barbero, M.I. (1996). Detección de funcionamiento diferencial de los ítems mediante análisis de residuales: una aplicación de la TRI. *Psicothema, 8*, 1, 173-180.

Prieto, P.; Barbero, M.I. y San Luis, C. (1997). Identification of nonuniform differential item functioning: A comparison of Mantel-Haenszel and Item Response Theory Analysis Procedures*. Educational and Psycho - logical Measurement, 57*, 4, 559-568.

Raju,N.S. (1988).The area between two item characteristic curves.*Psy - chometrika,53*, 492-502

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

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

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

Uttaro, T. y Millsap, R.E. (1994). Factors Influencing the Mantel-Haenszel Procedure in the Detection of Differential Item Functioning. *Ap - plied Psychological Measurement, 18*, 1, 15-25.

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

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