

## Procedimiento para la detección del FDI tanto en ítems politómicos como dicotómicos

M<sup>a</sup> Isabel Barbero García, Pedro Prieto y Concepción San Luis  
Universidad Nacional de Educación a Distancia

En el presente trabajo se expone una variación del procedimiento de detección del funcionamiento diferencial de los ítems (FDI), que dentro del marco de la TRI, y siguiendo la propuesta de Linn y Harnisch (1981) ha sido utilizado por Barbero, Sánchez Bruno, Prieto y San Luis (1995), Prieto y Barbero (1996), Prieto, Barbero y San Luis (1997), Barbero y Prieto (1997), entre otros. El objetivo de esta variación es el de solucionar el problema que se puede plantear con aquellos ítems que de entrada no presentan un buen ajuste en el grupo de referencia y sin embargo se aceptan como ítems que no presentan FDI en base a un buen ajuste con las respuestas de los sujetos en el grupo focal, cuando quizás la conclusión debería ser que si el ítem se ajusta correctamente a las respuestas de uno de los grupos pero no a las del otro, estamos en presencia de un ítem con FDI entre los grupos. Básicamente el procedimiento consiste en determinar el ajuste de las respuestas de los sujetos en el grupo focal, para luego comprobar el ajuste de los mismos datos pero con los parámetros previamente estimados en el grupo de referencia. La diferencia entre los dos índices de ajuste obtenidos se compara con un valor crítico obtenido mediante un procedimiento de simulación. En este estudio se ha aplicado el procedimiento a ítems de respuesta politómica, aunque es igualmente válido en ítems con respuesta dicotómica. Las condiciones manipuladas fueron tamaño y tipo del FDI, porcentaje de ítems con FDI y nivel de significación. Los resultados indican que el procedimiento propuesto es bastante eficaz en la detección del FDI, especialmente en el caso de FDI no uniforme y mixto, siendo la tasa de falsos positivos bastante baja, sobre todo cuando se compara con otros estudios de características similares.

*A procedure for detecting DIF in polythomous and dichotomous items.* In the present paper a modification of a DIF detection procedure is presented. The DIF procedure was proposed by Linn and Harnisch (1981) and used in applied settings by Barbero, Sánchez Bruno, Prieto and San Luis (1995), Prieto and Barbero (1996), Prieto, Barbero and San Luis (1997), Barbero and Prieto (1997), among others. The purpose of this modification was to be able to detect adequately the DIF when some items from the test don't fit to the data in the reference group. Basically, the procedure consists in several steps: first, to check the fit of the model to the subject responses in the focal group. Second, to estimate the item parameters in the reference group and to check the fit of these parameters in the focal group and, lastly, to calculate the difference between both fit indices in each item and to compare with a critical value obtained through a simulation procedure. This procedure can be used with both polythomous and dichotomous items. In this stimulation study, the procedure was applied to polythomous items and four factors were manipulated: percent of DIF presented in each item, type of DIF, percent of the items with DIF and significance level. The results indicates that the proposed procedure is effective in the detection of DIF, mainly when the DIF is non-uniform and mixed. Likewise, the rates of false alarms was very low, specially if it is compared with the rates from other simulation studies with similar features.

A lo largo de los últimos años se han desarrollado una gran variedad de procedimientos para la identificación y evaluación del

Funcionamiento Diferencial de los Ítems (FDI) tanto desde la perspectiva de la Teoría Clásica de los Tests (TCT), como desde la Teoría de Respuesta al Ítem (TRI). Aunque la mayoría de estos procedimientos se desarrollaron para ítems de respuesta dicotómica por ser el formato de respuesta característico de los tests de aptitudes, la búsqueda de otro tipo de formatos que dieran cuenta de la ejecución de los sujetos en los tests de aptitudes, junto a la necesidad de extender los estudios acerca del FDI a otro tipo de tests (actitudes, intereses, personalidad, etc.) en los que es habitual otro

tipo de formato de respuesta, ha motivado que en los últimos años aparezcan nuevas propuestas para la evaluación del FDI, o extensiones de los procedimientos ya existentes ajustados al nuevo tipo de formato. Así, Zwick, et al. (1992) proponen dos extensiones del estadístico Mantel-Haenszel: el procedimiento Mantel-Haenszel generalizado y el procedimiento Mantel-Haenszel politómico y, dentro del marco de la TRI, se proponen generalizaciones del estadístico  $\chi^2$  de Lord y de las medidas exactas de área de Raju propuestas por Cohen, Kim y Baker (1993) para el caso en que los ítems se ajusten al modelo de respuesta graduada.

La importancia del problema del FDI ha sido reconocida en España, prueba de ello son los numerosos trabajos publicados en torno al tema (Barbero, Sánchez Bruno, Prieto y San Luís, 1995; Barbero y Prieto, 1997; Fidalgo, 1994; Fidalgo y Paz, 1995; Fidalgo, Mellenbergh y Muñiz, 1998; Gómez y Navas, 1996, 1998; Hidalgo y López-Pina, 1997; Navas, 1994; Padilla, Pérez-Meléndez y González, 1998; Prieto y Barbero, 1996; Prieto, Barbero y San Luís, 1997 entre otros). En muchos de ellos se han hecho estudios comparativos para evaluar la precisión de distintos procedimientos para la identificación de ítems que presenten FDI y su comportamiento a la hora de detectar falsos positivos. Los resultados obtenidos parecen apuntar que uno de los más prometedores es el de Mantel-Haenszel propuesto por Holland y Thayer (1988), junto a las extensiones propuestas por Zwick et al. (1992) y entre los basados en los principios de la TRI merecen destacar las medidas exactas de área de Raju, el estadístico  $\chi^2$  de Lord (junto a las extensiones propuestas por Cohen, Kim y Baker, 1993) y el procedimiento basado en el análisis de residuales propuesto por Linn y Harnisch (1981).

En el presente estudio se propone una variación del método basado en el análisis de residuales utilizado por nuestro equipo en estudios anteriores, variación que puede ser utilizada tanto si las respuestas a los ítems son dicotómicas como si son politómicas.

Para llevar a cabo los primeros trabajos se desarrolló un programa al que denominamos GENESTE (San Luis, Prieto, Barbero y Sánchez-Bruno, 1995) que estaba formado por varios módulos y permitía llevar a cabo estudios de simulación mediante la generación de matrices de datos (GEN), estimación de parámetros (EST) y evaluación del ajuste (E). Para la primera y tercera fases se desarrolló un simulador de datos y un programa para el análisis de residuales que permitía evaluar la precisión de las estimaciones que se llevaban a cabo, la segunda fase se llevaba a cabo mediante el programa BILOG.

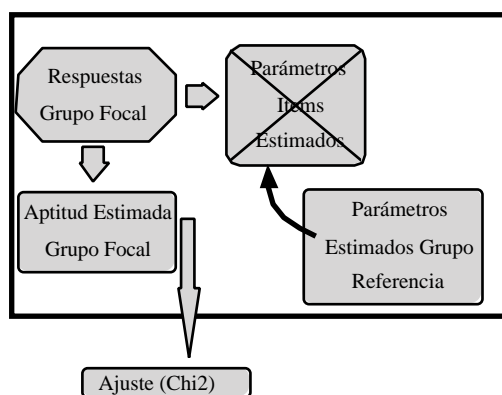


Gráfico 1. FDI mediante el análisis de residuales

El procedimiento consiste en determinar, en primer lugar, el modelo de la TRI que mejor se ajusta a los datos en un grupo de referencia y, a continuación, estimar un índice global de los residuos estandarizados obtenidos al evaluar el grado de ajuste del grupo focal con respecto a la curva característica del grupo de referencia, tal y como se esquematiza en el gráfico 1.

Como puede observarse en el esquema, se estima el nivel de habilidad de los sujetos en el grupo focal utilizando los parámetros de los ítems estimados en el grupo de referencia y, posteriormente, se evalúa el ajuste entre la distribución de las puntuaciones empíricas y las estimadas.

El módulo de análisis de residuales del programa GENESTE ofrece una serie de índices globales: proporción de residuos estandarizados menores o iguales que 1,96 (para un  $\alpha = 0,05$ ), residuo medio, residuo medio absoluto,  $\chi^2$  (Wright y Panchapakesan, 1969).

Según el procedimiento propuesto, se considera que un ítem no presenta FDI si el nivel de ajuste de los datos del grupo focal a las curvas características obtenidas en el grupo de referencia es adecuado. Si aplicamos este criterio de forma mecánica, puede suceder que un ítem cuyos parámetros no hubiesen sido estimados adecuadamente en el grupo de referencia fuera al final aceptado como un ítem que no presenta FDI en base a un buen ajuste con las respuestas de los sujetos del grupo focal, cuando quizás la conclusión debería ser que si el ítem se ajusta correctamente con las respuestas de uno de los grupos pero no con las del otro, estamos en presencia de un ítem con FDI entre los grupos.

Para tratar de dar solución a este problema se propone una modificación del procedimiento original que consiste en llevar a cabo las siguientes fases:

- Estimar los parámetros de los ítems tanto en el grupo focal como en el de referencia.
- Evaluar el ajuste del modelo en el grupo focal, mediante  $\chi^2$ .
- Evaluar el grado de ajuste del grupo focal, mediante  $\chi^2$ , con respecto a la curva característica del grupo de referencia; es decir, utilizando los parámetros de los ítems estimados en el grupo de referencia, previa estandarización de los mismos.
- Estimar para cada ítem las diferencias obtenidas entre los dos valores  $\chi^2$  obtenidos en ambas evaluaciones.
- Comparar las diferencias encontradas con un valor  $\chi^2$  crítico obtenido mediante simulación.

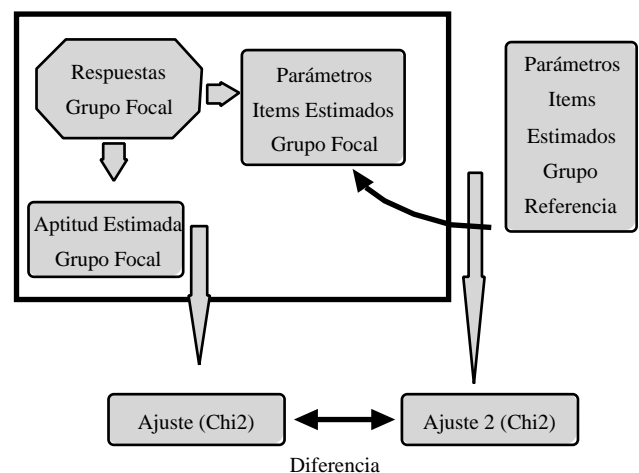


Gráfico 2. Modificación del procedimiento de análisis de residuales

Para la obtención de los valores críticos se hace lo siguiente: A partir de los parámetros obtenidos en el grupo de referencia y de los valores de habilidad  $\theta$  estimados en ese mismo grupo, se generan las respuestas aleatorias de distintas muestras de sujetos. Es de esperar que en estas muestras generadas a partir de los mismos parámetros, no encontremos ítems que presenten FDI. A estas muestras se les aplica el procedimiento de detección de FDI antes propuesto, obteniéndose así una distribución de diferencias de valores  $\chi^2$ . Como valor de comparación, valor crítico, se escoge el valor de la distribución de diferencias  $\chi^2$  que deje por debajo de sí el  $1-\alpha$  de la distribución. En caso de que el valor encontrado con los datos reales supere ese valor crítico se puede considerar que el ítem en cuestión presenta FDI.

La idea de generar distribuciones de muestras simuladas bajo la hipótesis de ausencia de FDI ha sido propuesta anteriormente por autores como Rogers y Hambleton (1989).

Para comprobar la bondad del método propuesto se ha diseñado un estudio de simulación en el que se le evalúa bajo diferentes condiciones: cuantía y tipo de FDI, así como porcentaje de ítems con FDI. El modelo de respuesta al ítem seleccionado ha sido el modelo de respuesta graduada de Samejima (1969), para lo cual se adaptó el programa GENESTE de manera que pudiera llevar a cabo los análisis necesarios; no obstante, el procedimiento propuesto, tal y como se ha comentado, puede ser utilizado con ítems de respuesta dicotómica y politómica.

Diseño:

- Para la evaluación del procedimiento de detección del FDI en ítems politómicos se generó una muestra aleatoria de 500 sujetos con distribución normal  $N(0,1)$ .

- Se generó asimismo un conjunto de 50 ítems con cinco categorías de respuesta.

- Los parámetros de discriminación de los ítems se generaron aleatoriamente en el intervalo 0,5 - 2.

- Los parámetros de dificultad se generaron con valores comprendidos entre -2,5 y 2,5.

En la tabla 1 se recogen los valores de los parámetros de los ítems obtenidos para el grupo de referencia

Para la creación del efecto del FDI, se ha seguido el procedimiento utilizado por Hidalgo y López (1997), estableciendo las mismas condiciones en las que se manipularon la cantidad de ítems con FDI en el test (10, 15 y 20 ítems con FDI), el tipo de FDI (uniforme, no uniforme y mixto) y la cuantía del FDI ( $d_{r-f} = 0,4$  y  $d_{r-f} = 0,7$ ).

El FDI uniforme se estableció haciendo que esas diferencias fueran entre los parámetros de dificultad de los límites de las categorías.

El FDI no uniforme se estableció de manera que las diferencias fueran entre los parámetros de discriminación de los grupos a comparar.

El FDI mixto se estableció haciendo que las diferencias fueran entre ambos parámetros.

Las condiciones fueron las siguientes:

$F_0$ : No FDI

$F_1$ : 10 ítems sesgados con diferencias entre los parámetros de 0,4

$F_2$ : 10 ítems sesgados con diferencias entre los parámetros de 0,7

$F_3$ : 15 ítems sesgados con diferencias entre los parámetros de 0,4

$F_4$ : 15 ítems sesgados con diferencias entre los parámetros de 0,7

$F_5$ : 20 ítems sesgados con diferencias entre los parámetros de 0,4

$F_6$ : 20 ítems sesgados con diferencias entre los parámetros de 0,7.

En cada una de las condiciones el tipo de FDI generado fue uniforme, no uniforme y mixto, así en:

- $F_1$  y  $F_2$  se introdujeron 4 ítems con FDI uniforme, 3 con FDI no uniforme y 3 con FDI mixto.

- $F_3$  y  $F_4$  se introdujeron 5 ítems con FDI uniforme, 5 con FDI no uniforme y 5 con FDI mixto.

- $F_5$  y  $F_6$  se introdujeron 7 ítems con FDI uniforme, 7 con FDI no uniforme y 6 con FDI mixto

Tabla 1  
Parámetros verdaderos de los ítems del grupo de referencia

Item	a	b1	b2	b3	b4
1	1.4726	-0.2286	1.4494	2.1857	2.4905
2	0.9685	-1.0909	-0.7831	2.0114	2.4405
3	0.8966	-1.9386	1.7849	2.0187	2.4007
4	0.9161	-1.9388	0.7777	1.4346	2.4499
5	0.6511	-2.3185	-1.1345	-0.2641	2.1548
6	0.6079	-1.3105	0.9443	1.2933	2.0710
7	1.2303	-2.0386	-1.5417	0.4655	2.4300
8	1.2343	-1.9569	-1.2147	0.9726	1.7257
9	0.9200	-1.0280	0.0481	0.8579	1.9327
10	1.5510	-0.5775	0.4045	1.4274	2.4652
11	1.7775	-1.0733	0.1864	1.4881	2.4994
12	0.5345	-0.5926	-0.0327	1.5247	2.3303
13	1.5997	-2.2397	0.3491	1.4247	2.4916
14	1.6330	-0.7239	1.4769	2.3949	2.9506
15	1.2916	-1.5260	-0.4922	0.8046	2.2308
16	0.8514	-2.4282	-0.3229	1.3611	2.4125
17	1.2206	-0.1050	0.2847	1.3169	2.4756
18	1.7323	-1.7009	0.3292	1.3354	2.3784
19	1.6344	-2.1911	-1.0933	1.3202	2.4358
20	1.7319	-2.1113	0.2175	1.0339	2.1945
21	0.8378	-1.9955	-1.3804	-0.7211	0.3611
22	1.2074	-1.4242	0.8851	1.5622	2.0920
23	1.7484	-1.1240	-0.5216	1.8704	2.6946
24	1.7942	-1.1258	0.9339	1.6732	2.7884
25	1.7690	-2.1371	-0.8606	1.6503	2.4563
26	1.2325	-1.2886	0.8924	1.9042	2.4610
27	1.4778	-0.4812	1.1438	1.7910	2.4069
28	1.0133	-0.4389	0.0339	0.7272	1.9885
29	1.4602	-2.3353	1.6315	2.0543	2.8899
30	1.7383	-1.9655	0.2720	1.6970	2.4389
31	1.5646	-0.8501	-0.0100	0.9060	1.5296
32	1.1484	-2.1712	-0.5126	0.8451	1.7101
33	1.3127	-0.7700	1.0228	1.9617	2.3570
34	1.2414	-2.2184	0.7588	1.6219	2.2074
35	1.1400	-0.6181	1.2158	1.9102	2.4169
36	0.6590	-2.3576	-1.5524	1.2777	2.4732
37	0.7417	-0.6282	0.1171	1.1209	2.2457
38	1.7955	-0.9908	0.4933	1.5840	2.0143
39	1.8504	-0.9512	0.3466	1.2271	2.2287
40	1.2649	-0.3678	0.0134	0.3623	1.2900
41	1.1218	-0.7616	1.5365	2.0096	2.4720
42	1.6268	-0.5738	0.3018	1.4756	2.4820
43	1.0469	-0.9406	-0.9326	1.5088	2.0725
44	1.9767	-0.3312	0.3528	1.8663	2.4652
45	1.9518	-2.2960	-1.3364	-0.8329	0.6584
46	1.4151	-1.6943	1.0916	1.8428	2.4697
47	0.8148	-2.4841	0.9092	1.9509	2.4870
48	1.5488	-1.7856	0.2608	1.9209	2.4431
49	0.8231	-1.1286	1.2966	1.9389	2.4917
50	1.4683	-1.3556	0.4938	1.4979	2.4993

*Tabla 2*  
Recubrimiento de los parámetros

Grupo	θ		a		b1		b2		b3		b4		Modelo Residuos
	r	rmse	r	rmse	r	rmse	r	rmse	r	rmse	r	rmse	
refe	.976	0.24	.968	0.17	.954	0.23	.994	0.10	.964	0.25	.811	0.65	0.9405
sim	.976	0.24	.923	0.11	.967	0.21	.989	0.14	.971	0.21	.941	0.47	0.9400
sim1	.980	0.21	.943	0.10	.973	0.18	.990	0.14	.965	0.23	.930	0.39	0.9440
sim2	.978	0.22	.912	0.12	.980	0.17	.991	0.12	.981	0.16	.944	0.49	0.9335
f0	.978	0.24	.949	0.18	.982	0.16	.990	0.13	.953	0.24	.794	0.61	0.9530
f1	.979	0.24	.954	0.20	.959	0.25	.996	0.10	.969	0.23	.887	0.60	0.9435
f2	.979	0.25	.952	0.19	.954	0.30	.991	0.13	.975	0.21	.870	0.62	0.9490
f3	.981	0.24	.962	0.21	.961	0.28	.989	0.13	.972	0.19	.861	0.55	0.9450
f4	.982	0.24	.955	0.25	.949	0.31	.991	0.12	.977	0.22	.869	0.57	0.9510
f5	.979	0.25	.963	0.20	.940	0.36	.990	0.14	.965	0.27	.874	0.63	0.9335
f6	.982	0.25	.960	0.29	.945	0.43	.987	0.17	.968	0.24	.888	0.65	0.9455

Generación de datos

Además de los datos correspondientes al grupo de referencia y tres muestras de línea base, para cada una de las condiciones se generó una matriz con las respuestas simuladas de 500 sujetos a 50 ítems. Para llevar a cabo la generación de los datos se utilizó el programa GENESTE que habíamos desarrollado para ítems dicotómicos y que ha sido adaptado para la generación de datos siguiendo el modelo de respuesta graduada de Samejima.

En cada caso se estudió el recubrimiento de los parámetros a través de dos vías:

- a) Mediante la raíz del error cuadrático medio
- b) A través de la correlación entre los parámetros verdaderos y los estimados (véase Tabla 2).

Como puede apreciarse, el nivel de precisión en la estimación de los parámetros es bastante alto en todas las condiciones. También se calculó en cada ocasión la proporción de residuos estandarizados con valores absolutos mayores que 2. Se puede observar en la misma tabla (última columna) que en todos los casos los datos presentaban un buen ajuste al modelo elegido.

Obtención de los puntos críticos

Como paso previo para poder aplicar el procedimiento propuesto para la evaluación del FDI, se generaron a partir de los parámetros de los ítems obtenidos en el grupo de referencia y de los valores de θ de

los sujetos de ese mismo grupo, tres matrices diferentes de datos. Cada una de estas muestras se tomo como grupo focal y de acuerdo con el procedimiento planteado se estimó, para cada ítem el valor de la diferencia entre el  $\chi^2$  obtenido al evaluar el ajuste del modelo en el grupo focal con sus propios parámetros y el  $\chi^2$  obtenido al evaluar el ajuste del modelo en el grupo focal pero utilizando los parámetros estimados en el grupo de referencia. Dado que son tres muestras de 50 ítems cada una, nos encontramos con 150 diferencias de valores  $\chi^2$ .

*Tabla 3*  
Proporciones de ítems correctamente detectados y falsos positivos

Proporción ICD (.975)				
Grupo	Tipo de FDI			total
	uniforme	no uniforme	mixto	
focal 1	0.25	1.00	1.00	0.70
focal 2	1.00	1.00	1.00	1.00
focal 3	0.20	0.40	0.60	0.40
focal 4	0.80	0.80	1.00	0.87
focal 5	0.28	0.28	0.66	0.30
focal 6	0.71	0.57	1.00	0.75
total	0.53	0.60	0.78	0.64
Proporción ICD (.95)				
Grupo	Tipo de FDI			Total
	uniforme	no uniforme	mixto	
focal 1	0.50	1.00	1.00	0.80
focal 2	1.00	1.00	1.00	1.00
focal 3	0.20	0.60	0.80	0.53
focal 4	0.80	0.80	1.00	0.87
focal 5	0.28	0.42	0.50	0.40
focal 6	0.71	0.71	1.00	0.80
total	0.56	0.70	0.86	0.71
Falsos Positivos				
		97,5%		95%
focal 0		0.020		0.020
focal 1		0.000		0.025
focal 2		0.000		0.025
focal 3		0.000		0.000
focal 4		0.028		0.028
focal 5		0.000		0.000
focal 6		0.000		0.033
total		0.004		0.019

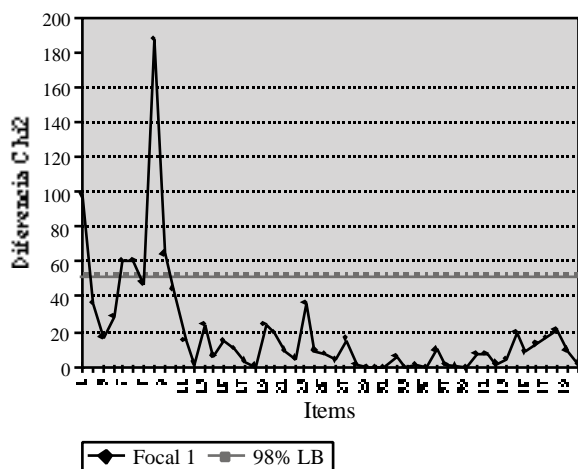


Gráfico 3. Resultados en el grupo focal, condición F1

La distribución de frecuencias acumuladas de estos valores puso de manifiesto que el valor de la diferencia correspondiente al percentil 97,5 fue 43,51 y para el percentil 95 fue 36,50. Por lo tanto, estos son los valores que se tomaron como valores críticos para  $\alpha = 0,025$  y  $0,05$  respectivamente.

Una vez establecidos los valores críticos, el siguiente paso consistió en aplicar el procedimiento propuesto a cada una de las condiciones y comparar, para cada ítem, el valor de la diferencia entre los dos  $\chi^2$  obtenidos y los valores críticos.

En el gráfico 3 se presenta el resultado obtenido al hacer estas comparaciones con el grupo focal de la condición  $F_1$  (10 ítems con FDI 0,4), para un valor de  $\alpha = 0,020$ . Se puede observar que únicamente 5 ítems superan el punto crítico correspondiente al  $\alpha$  utilizado, por lo que podemos considerarlos como ítems que presentan FDI entre los grupos a comparar.

#### Tasa de ítems correctamente detectados

En la tabla 3 se presenta la tasa de ítems correctamente detectados en las diferentes condiciones y para distintos valores de  $\alpha$ . Como puede apreciarse, la proporción total de ítems correctamente detectados fue de 0,64 para  $\alpha = 0,025$  y de 0,71 para  $\alpha = 0,05$ .

La tasa parece aumentar en el caso de ítems con FDI mixto. Asimismo se puede observar cómo la tasa de aciertos aumenta, como era de esperar, a medida que aumenta la cuantía de FDI aunque, por otra parte, y de forma inesperada, parece que la misma se reduce a medida que aumenta el número de ítems con FDI.

En el estudio de Hidalgo y López (1997), en el que utilizaron el  $\chi^2$  de Lord y la Z(ACS) de Cohen y colaboradores como estadísticos para detectar el FDI, la tasa de ítems correctamente identificados fue de 0,73 y 0,78 respectivamente para un  $\alpha = 0,05$ . Sin

embargo en el caso de ítems con FDI no uniforme la proporción bajaba a un 0,22 y 0,42 respectivamente, mientras que con el procedimiento que proponemos esa tasa alcanza un nivel de 0,70 para el mismo valor de  $\alpha$ .

#### Tasa de falsos positivos

En este sentido los resultados son bastante alentadores, ya que en ningún caso se sobrepasó una tasa del 3% mientras que en otros estudios se llegó a tasas cercanas al 50%. En todo caso, las tasas de falsos positivos son sensiblemente inferiores a las encontradas por nosotros en anteriores trabajos, previos a la modificación propuesta al procedimiento original y en los que se utilizaban modelos de respuesta dicotómica.

#### Conclusiones

La importancia del problema del FDI se ha puesto de manifiesto en el número de trabajos que se están llevando a cabo y en la cantidad de procedimientos que se han desarrollado para su detección y evaluación. Muchos de ellos se mostraron eficaces para la detección de FDI uniforme, aunque no tanto cuando el FDI era no uniforme y mixto. Los trabajos realizados con el programa GENESTE, al permitir mediante representación gráfica la visualización de los residuos estandarizados, permitió mejorar la detección de ítems con FDI no uniforme. El procedimiento que presentamos en este trabajo aumenta la tasa de identificación de ítems con FDI mixto y reduce considerablemente la tasa de falsos positivos.

En este trabajo se ha aplicado el procedimiento propuesto para la evaluación del FDI globalmente, en este momento se está desarrollando un nuevo trabajo en el que se está probando la eficacia del procedimiento para evaluar el FDI a través de los límites de las categorías.

#### Referencias

- Barbero, M.I., Sánchez-Bruno, J.A., Prieto, P. y San Luis, C. (1995). *DIF uniforme versus no uniforme: análisis mediante el programa GENESTE*. IV Simposio Nacional de Metodología de las Ciencias del Comportamiento. Murcia
- Barbero, M.I. y Prieto, P. (1997). Evaluación del rendimiento en Ciencias de los niños y niñas de 13 años de las distintas Comunidades Autónomas: Impacto o sesgo. *Psicothema*, 9 (2), 433-440.
- Cohen, A.S., Kim, S.H. y Baker, E. (1993) Detection of Differential Item Functioning in the Graded Response Model. *Applied Psychological Measurement*, 17, 335-350.
- 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.
- Fidalgo, A.M. y Paz, M.D. (1995). Modelos lineales logarítmicos y funcionamiento diferencial de los ítems. *Anuario de Psicología*, 64, 57-66.
- Fidalgo, A.M., Mellenberg, G. y Muñoz, 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. y Navas, M.J. (1996). Detección de sesgo mediante regresión logística: purificación paso a paso de la habilidad. *Psicológica*, 17 (2), 397-411.
- Gómez, J. y Navas, M.J. (1998). Impacto y funcionamiento diferencial de los ítems respecto al género en una prueba de aptitud numérica. *Psicothema*, 10 (3), 685-696.
- Hidalgo, M.D. y López, J.A. (1997). Evaluación del funcionamiento diferencial en ítems politómicos mediante el estadístico de Lord y las medidas de área. *Psicológica*, 18, 69-92.
- Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In Wainer, H. & Braun, H.I. (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.
- Linn, R.L. y Harsnich, D.L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement*, 18, 109-118.
- Navas, M.J. (1994). Utilización del análisis factorial y medidas de área como métodos en la detección de sesgo. *Psicothema*, 6 (3) 493-501.
- Padilla, J.L., Pérez-Meléndez, C. y González, A. (1998). La explicación del sesgo en los ítems de rendimiento. *Psicothema*, 10 (2) 481-490.
- Prieto, P. y Barbero, M.I. (1996). Detección del 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 DIF: A comparison of Mantel-Haenszel and IRT analysis procedure. *Educational and Psychological Measurement*, 57(4), 559-568.
- Rogers, H.J. y Hambleton, R.K. (1989). Evaluation of computer simulated baseline statistics for use in item bias studies. *Educational and Psychological Measurement*, 49, 355-369.
- Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement* 17.
- San Luis, C., Prieto, P., Barbero, M. I. y Sánchez-Bruno, J.A. (1995). GENESTE: Un programa de control para TRI. *Psicológica*, 16, 297-304.
- Wright, B.D. y Panchapakesan, N. (1969). A procedure for sample free item analysis. *Educational and Psychological Measurement*, 29, 23-48.
- Zwick, R., Donoghue, J., Grima, A., Holland, P.W., Thayer, D., Thomas, N. y Wingersky, M. (1992) *Differential item functioning analysis for new models of assessment*. Paper presentado en la Annual Convention of the National Council on Measurement in Education. San Francisco.