Psicothema
2000. Vol. 12, nº 3, pp. 495-502

ISSN 0214 - 9915 CODEN PSOTEG

Influencia de la igualación iterativa en la detección del funcionamiento diferencial del ítem mediante las medidas de área de Raju y el estadístico de Lord

Rosa Mª Núñez Núñez, Mª Dolores Hidalgo Montesinos y José Antonio López Pina
Universidad de Murcia

Con objeto de mejorar la identificación correcta de los ítems de un test que presentan funcionamiento diferencial (DIF), se han propuesto diversos procesos de purificación del test. El presente estudio emplea un procedimiento de purificación bietápico (Hidalgo y López, 1997b) para mejorar la detección del DIF con el estadístico de Lord y las medidas de área de Raju. Las condiciones manipuladas fueron: tamaño muestral (250, 500 y 1.000 sujetos), cantidad de DIF (0.4 y 0.6), tipo de DIF (uniforme, no uniforme y mixto) y porcentaje de ítems con DIF en el test (10% y 30%). Los resultados indican que el procedimiento bietápico reduce las tasas de falsos positivos de dichas medidas.

Influence of iterative linking on differential item functioning detection using Raju's area measures and Lord's statistic. In order to improve the correct identification of test items which show differential functioning (DIF), several test purification procedures have been proposed. The present study uses a bietapic purification procedure (Hidalgo & López, 1997b) to improve DIF detection with Lord's χ^2 statistic and Raju's area measures. Sample size (250, 500 and 1.000 subjects), amount of DIF (0.4 and 0.6), types of DIF —uniform, non-uniform and mixed— and percent of DIF test items (10% and 30%) were manipulated. The results show that the bietapic procedure reduces the false positive rates.

En la evaluación de variables psicológicas y educativas, los términos sesgo del ítem/test y funcionamiento diferencial del ítem/test (DIF/DTF) hacen referencia a la inconsistencia de las propiedades psicométricas del ítem/test a través de distintas muestras pertenecientes a una misma población. Dentro del marco de la Teoría de Respuesta al Item (TRI), un ítem funciona diferencialmente si individuos pertenecientes a grupos distintos pero con el mismo nivel de habilidad, tienen diferentes probabilidades de responder correctamente al ítem, lo cual implica que los parámetros del ítem difieren entre los grupos, infringiéndose el supuesto de invarianza de los parámetros del ítem. Gráficamente, un ítem funciona diferencialmente si su curva característica (CCI) no es la misma para todos los grupos.

Normalmente, en la evaluación del DIF, el grupo objeto de análisis se denomina grupo focal (F) y el grupo que sirve como criterio de comparación se conoce como grupo de referencia (R). Mellenbergh (1982) distingue los siguientes tipos de DIF: 1) DIF uniforme, cuando la probabilidad de responder correctamente a un ítem es mayor, a lo largo de toda la escala de habilidad, en un grupo (p.e. grupo de R) que en otro (p.e. grupo F), porque sólo hay diferencias en el parámetro de dificultad del ítem ($b_{jF} \neq b_{jR}$) y 2) DIF no uniforme, cuando la diferencia de las probabilidades de responder correctamente al ítem en los dos grupos no es constante a lo largo del continuo de habilidad, porque las diferencias aparecen, o bien en el parámetro de discriminación ($a_{jF} \neq a_{jR}$) (DIF no uniforme propiamente dicho, o DIF no uniforme simétrico), o bien en todos los parámetros del ítem ($a_{jF} \neq a_{jR}$, $b_{jF} \neq b_{jR}$ y $c_{jF} \neq c_{jR}$) (DIF mixto o no uniforme asimétrico).

En la actualidad se dispone de una amplia gama de técnicas para evaluar el funcionamiento diferencial del ítem en formatos de respuesta dicotómica (Gómez e Hidalgo, 1997; Millsap y Everson, 1993). Estas técnicas se pueden clasificar como (Millsap y Everson, 1993): a) de invarianza condicional observada, que utilizan las puntuaciones observadas en el test como variable de equiparación y b) de invarianza condicional no observada, donde la variable de equiparación es una estimación de la habilidad, por ejemplo desde la TRI.

La identificación de ítems con DIF, independientemente de la técnica utilizada, se ve afectada por la presencia en el test de otros ítems con DIF, siendo el principal problema un incremento del número de falsas identificaciones de ítems con DIF. Las técnicas de detección del DIF que utilizan la puntuación total en el test como un estimador de la habilidad han implementado diversos procesos de purificación del test (Fidalgo, Mellenbergh y Muñiz, 1998; Fidalgo y Paz, 1995; Gómez y Navas, 1996; Holland y Thayer, 1988; Kok, Mellenbergh y Van der Flier, 1985; Van der Flier, Mellenbergh, Adèr y Wijn, 1984), con la finalidad de paliar dichos problemas. Estos procedimientos son el punto de referencia de los elaborados en el marco concreto de la TRI (Candell y Drasgow, 1988; Hidalgo y López, 1997b; Lautenschlager, Flaherty y Park, 1994; Lord 1980; Miller y Oshima, 1992; Park y Lautenschlager,

Correspondencia: Rosa Mª Núñez Núñez
Facultad de Psicología
Universidad de Murcia
30080 Murcia (Spain)
E-mail: romanu@fcu.um.es

1990). El proceso de detección del DIF, para la mayoría de los procedimientos desarrollados desde la TRI, implica en primer lugar la estimación de los parámetros del ítem en cada grupo por separado (supuesto que los grupos pertenecen a la misma población). En segundo lugar, y una vez estimados los parámetros, se equiparan o igualan las métricas, es decir, para poder comparar dichos parámetros estos deben estar en la misma escala de medida, requisito que se logra al emplear un *método de igualación*. El proceso de igualación requiere calcular las constantes de una función que permite igualar los parámetros de las curvas de respuesta al ítem en cada grupo. Por último, se evalúa el DIF utilizando aproximaciones tales como el cálculo del área entre las CCIs del grupo focal y de referencia, las medidas de área exacta de Raju (1990) o la comparación de parámetros de ítems (Lord, 1980). En general, los procedimientos de purificación del test en TRI se centran en la reestimación de los parámetros de ítems y/o en la reigualación de las métricas una vez eliminados del test los ítems con DIF.

En Hidalgo y López (1997a) se pone de manifiesto, que en tests con más de un 30% de ítems con DIF, el estadístico de Lord (1980) y las medidas de área de Raju (1990) obtienen una elevada tasa de falsos positivos. Una solución a este problema sería la utilización de algún procedimiento de purificación del test. El presente trabajo estudia los efectos de la purificación del test en la detección del DIF mediante las medidas de área de Raju y el estadístico de Lord bajo el modelo logístico de 2-p. En concreto, se implementa un proceso de purificación bietápico (Hidalgo y López, 1997b) que anteriormente se ha aplicado en la detección del DIF uniforme en ítems politómicos.

Procedimientos de evaluación del DIF

Estadístico de Lord

El estadístico de Lord (1980) es un método de evaluación del DIF basado en la comparación de los parámetros de los ítems. Con este objetivo se define la hipótesis nula de igualdad de parámetros del ítem en dos grupos de examinados (grupo F y grupo R), H_0 : $a_{jF} = a_{jR}$, $b_{jF} = b_{jR}$, $c_{jF} = c_{jR}$, donde a_{jF} y a_{jR} son los parámetros de discriminación, b_{jF} y b_{jR} y son los parámetros de dificultad, y c_{jF} y c_{jR} y son los parámetros de pseudo-azar del ítem en los grupos R y F. Un ítem no presenta DIF si tiene los mismos parámetros en los dos grupos, comprobado previamente, el ajuste del modelo de TRI.

El estadístico de Lord es un valor resultante de una operación vectorial:

$$x^2 = \left(\xi_{jR} - \xi_{jF}\right)' \Sigma_j^{-1} \left(\xi_{jR} - \xi_{F}\right)$$

$$\xi_{jR} = \left(\hat{a}_{jR} \hat{b}_{jR}\right)'$$

$$\xi_{jF} = \left(\hat{a}_{jF} \hat{b}_{jF}\right)'$$

y Σ_j es la matriz de dispersión 2 x 2, tal que $\Sigma_j = \Sigma_{jR} + \Sigma_{jF}$, donde Σ_{jR} y Σ_{jF} son las matrices de varianza-covarianza de ξ_{jR} y ξ_{JF}, respectivamente.

El estadístico de Lord sigue una distribución χ^2 con grados de libertad igual al número de parámetros del modelo ajustado.

Medidas de área exactas de Raju

Raju (1988) aportó las medidas de área exactas con signo (SA) y sin signo (UA) como índices de detección y evaluación DIF, para cada modelo logístico de respuesta al ítem (de 1-p y 2-p) y para el modelo de 3-p, en función de la igualdad y/o desigualdad de los parámetros de discriminación y pseudo-azar, en dos grupos comparados. Cada una de estas medidas de área va acompañada de la distribución muestral asintótica (Raju, 1990), con lo cual, es posible estudiar si las diferencias en las CCIs de los grupos se deben a errores aleatorios de muestreo o son diferencias estadísticamente significativas, utilizando una prueba Z. Para el modelo de 2-p, estas medidas son:

$$SA = \hat{b}_{jF} - \hat{b}_{jR}, \text{ y}$$

$$UA = \left|\hat{b}_{jF} - \hat{b}_{jR}\right| \text{ si } \hat{a}_{jR} = \hat{a}_{jF}, \text{ o } UA = \left|H_j\right| \text{ en cualquier otrocaso},$$

donde,

$$H_j = \frac{2\left(\hat{a}_{jF} - \hat{a}_{jR}\right)}{D\hat{a}_{jF}\hat{a}_{jR}} \ln\left\{1 + \exp\left[\frac{D\hat{a}_{jF}\hat{a}_{jR}\left(\hat{b}_{jF} - \hat{b}_{jR}\right)}{\hat{a}_{jF} - \hat{a}_{jR}}\right]\right\} - \left(\hat{b}_{jF} - \hat{b}_{jR}\right)$$

La prueba estadística, Z_j, para se define como:

$$Z(SA)_j = \frac{\hat{b}_{jF} - \hat{b}_{jR}}{\left[Var\left(\hat{b}_{jF}\right) + Var\left(\hat{b}_{jR}\right)\right]^{1/2}}$$

Sin embargo, no se puede asumir que la medida de área sin signo, UA, se distribuya normalmente cuando $\hat{a}_{jR} \neq \hat{a}_{jF}$, por lo que Raju (1990) sugiere que sea la variable H_j la que se utilice para probar la significación de UA. Entonces, la prueba estadística se define como:

$$Z_j(H) = \frac{H_j}{\left[Var\left(H_j\right)\right]^{1/2}}$$

Procesos de purificación del test

El DIF se evalúa asumiendo que los restantes ítems del test no presentan DIF, circunstancia que raramente sucede en la práctica. Tal y como se ha comentado anteriormente, las propuestas para solucionar este problema implica utilizar algún procedimiento de purificación del test.

Lord (1980) elaboró un proceso de purificación del test basado en estimación iterativa de los parámetros de habilidad y de los ítems, conforme éstos últimos son identificados con DIF y eliminados del test. La estimación de la habilidad se realiza con todos los grupos implicados en la realización del test, mientras que los parámetros de los ítems se estiman en cada grupo por separado.

Candell y Drasgow (1988) incorporan al procedimiento anterior un método de igualación de métricas que simplifica los cálculos, ya que no reestiman los parámetros de los ítems y de la habilidad al término de cada iteración, sino que se centran en el cálcu-

lo de las constantes de igualación conforme son eliminados del test los ítems identificados con DIF al finalizar cada iteración. La ventaja de este procedimiento frente al de Lord (1980) es su facilidad de implementación dado que no requiere la reestimación iterativa de los parámetros de los ítems y de la habilidad. Cohen y Kim (1993) implementaron el proceso iterativo de Candell y Drasgow (1988). Trabajaron con datos simulados en dos tamaños muestrales (100 y 500 sujetos) y en dos tests de ítems dicotómicos (20 y 60 ítems); los parámetros de los ítems se ajustaban al modelo de 2-p. La proporción de ítems con DIF para ambos test fue del 10% y 20%; la cantidad de DIF inducido fue d=0.5 y d=1 para el DIF uniforme, d=-0.5 para el no uniforme y, para el tipo mixto d=-0.5 en el parámetro de discriminación y d=0.5 en el parámetro de dificultad. Para estimar los parámetros del modelo de respuesta al ítem emplearon el método de estimación marginal bayesiana y el método de estimación por máxima verosimilitud marginal, en situaciones de no impacto y de impacto; los autores también estudiaron el efecto del método de estimación en la detección del DIF. El procedimiento de purificación de Candell y Drasgow (1988) produjo poca variación en los porcentajes de falsos positivos y falsos negativos a través de las iteraciones dentro de cada una de las condiciones experimentales.

Otro procedimiento de purificación basado en el trabajo pionero de Lord (1980), fue desarrollado por Park y Lautenschlager (1990), denominado M-LTP (Modified-Lord Test Purification), también centrado en la reestimación continua de los parámetros de los ítems y de la habilidad. Sin embargo, este procedimiento mostró peores resultados en la identificación del DIF que el procedimiento propuesto por Candell y Drasgow (Park y Lautenschlager, 1990). Este procedimiento fue mejorado por el denominado ILAP (Iterative Linking and Ability Purification) (Lautenschlager, Flaherty y Park, 1994; Park y Lautenschlager, 1990), el cual combina los procedimientos de Lord (1980) y de Candell y Drasgow (1988). En este procedimiento se estiman los parámetros de la habilidad y de los ítems iterativamente, y se igualan las métricas de los grupos comparados iterativamente. El procedimiento ILAP es efectivo pero costoso computacionalmente, por lo que Miller y Oshima (1992) proponen un procedimiento bietápico que reduce a dos iteraciones el procedimiento ILAP y reestiman solamente los parámetros de los ítems. Miller y Oshima (1992) señalaron que este procedimiento bietápico presentaba mejores resultados cuando el número de ítems sesgados en el test fue elevado (20% o más) y la magnitud del DIF fue moderada (una diferencia inducida en el parámetro de dificultad de 0.35).

En Hidalgo y López (1997b) se proponen un proceso bietápico de purificación del test a partir de algunas modificaciones en los procedimientos de Candell y Drasgow (1988) y de Miller y Oshima (1992). Este procedimiento consiste en:

Etapa 1:

1. Estimación de los parámetros de los ítems para cada grupo por separado.
2. Igualación de métricas.
3. Cálculo de la medida de evaluación del DIF.
4. Eliminar del test los ítems identificados con DIF.

Etapa 2:

1. Igualación de las métricas de los ítems identificados sin DIF en el paso anterior.
2. Cálculo de la medida de evaluación del DIF para todos los ítems.

Este procedimiento ha sido probado en tests de respuesta politómica utilizando las extensiones del estadístico de Lord y las medidas de área de Raju. Los resultados demostraron que, el procedimiento bietápico de purificación fue efectivo, incrementó las tasas de identificaciones correctas y redujo las de falsos positivos para ambas medidas. Además, computacionalmente es menos costoso dado que no requiere la reestimación de los parámetros de ítems.

Método

En este trabajo se ha utilizado un test de 40 ítems de respuesta dicotómica que se ajustan a un modelo logístico de 2-p, los parámetros de discriminación y dificultad fueron tomados de Candell y Drasgow (1988). Con la finalidad de evaluar el posible efecto del tamaño muestral sobre la potencia de las dos medidas para detectar el DIF, se han considerado muestras de 250, 500 y 1000 sujetos tanto para el grupo de referencia como para el grupo focal. Para cada uno de los tamaños muestrales se generó una distribución de habilidad normal tipificada en el intervalo [-2.5, +2.5].

Antes de generar las respuestas de los sujetos, se incrementaron los valores iniciales de los parámetros de los ítems para los grupos focales, con dos cantidades de DIF, d=0.4 y d=0.6; para cada una de estas dos cantidades se barajaron dos porcentajes de ítems con DIF, un 10% y 30% de los ítems del test (tabla 1), provocando los tres tipos de DIF posibles. Así, se obtienen cuatro condiciones experimentales:

• condición 1: d=0.4 y el 10% de ítems con DIF, de los cuales los dos primeros ítems presentan DIF uniforme, el tercer ítem presenta DIF no uniforme, y el cuarto ítem DIF mixto.

• condición 2: d=0.4 y el 30% de ítems con DIF, de los cuales los cuatro primeros ítems presentan DIF uniforme, los ítems 5, 6, 7 y 8 presentan DIF no uniforme, y el 9, 10, 11 y 12 contienen DIF mixto.

• condición 3: d=0.6 y el 10% de ítems con DIF, de los cuales los ítems 1 y 2 presentan DIF uniforme, el ítem 3 presenta DIF no uniforme, y el ítem 4 DIF mixto.

• condición 4: d=0.6 y el 30% de ítems con DIF, de los cuales los ítems 1, 2, 3 y 4 presentan DIF uniforme, los ítems 5, 6, 7 y 8 presentan DIF no uniforme, y el 9, 10, 11 y 12 contienen DIF mixto.

Los parámetros de los ítems del test que no se manipularon, 36 ítems en las condiciones 1 y 3, y 28 ítems en las otras dos, conservaron los valores originales (tabla 2).

Tabla 1
Parámetros de los ítems en las 4 condiciones experimentales

Item	Grupo de Referencia		Grupo focal Condición 1		Grupo focal Condición 2		Grupo focal Condición 3		Grupo focal Condición 4	
1	-1.23	0.90	-0.83	0.90	-0.83	0.90	-0.63	0.90	-0.63	0.90
2	-0.18	0.94	0.22	0.94	0.22	0.94	0.42	0.94	0.42	0.94
3	-1.06	1.16	-1.06	1.56	-0.66	1.16	-1.06	1.76	-0.46	1.16
4	-0.92	0.88	-0.52	1.28	-0.52	0.88	-0.32	1.48	-0.32	0.88
5	-0.71	0.97			-0.71	1.37			-0.71	1.57
6	0.42	0.72			0.42	1.12			0.42	1.32
7	0.77	0.37			0.77	0.77			0.77	0.97
8	1.53	0.86			1.53	1.26			1.53	1.46
9	-0.06	1.05			-0.34	1.45			0.54	1.65
10	-0.54	1.07			-0.14	1.47			0.06	1.67
11	-0.02	1.04			0.38	1.44			0.58	1.64
12	-1.24	1.48			-0.84	1.88			-0.64	2.08

En total se producen 12 combinaciones: 3 tamaños muestrales x 2 cantidades de DIF x 2 porcentajes de ítems con DIF. Las respuestas a los ítems de los tres grupos de referencia y de los doce grupos focales fueron simuladas con el programa SIMULA v. 2 (Hidalgo y López, 1995). Para cada uno de ellos se generaron 50 réplicas con objeto de reducir los errores aleatorios del proceso de muestreo y dar estabilidad a los resultados.

A partir de estas matrices se estimaron los parámetros de los ítems. Las estimaciones fueron realizadas con el programa BILOG v. 3.04 (Mislevy y Bock, 1990) utilizando las opciones por defecto del mismo.

La igualación de parámetros se llevó a cabo con el programa EQUATE v. 2.0 (Baker, 1993), que ejecuta el método de las curvas características de Stocking y Lord (1983). Una vez igualadas las métricas se calcularon, para cada ítem del test, tanto el estadístico de Lord como las medidas de área, utilizando el programa IRTDIF (Kim y Cohen, 1992). A continuación, los ítems detectados con funcionamiento diferencial fueron eliminados del test. Se procedió de nuevo a obtener las constantes de igualación y se calcularon otra vez los estadísticos del DIF para todos los ítems del test.

Tabla 2
Parámetros del grupo focal en las distintas condiciones experimentales

Item	Grupo focal Condición 1 y 3		Grupo focal Condición 2 y 4	
5	-0.71	0.97		
6	0.42	0.72		
7	0.77	0.37		
8	1.53	0.86		
9	-0.06	1.05		
10	-0.54	1.07		
1	-0.02	1.04		
12	-1.24	1.48		
13	-0.34	0.91	-0.34	0.91
14	-1.15	1.09	-1.15	1.09
15	-1.42	1.17	-1.42	1.17
16	-0.86	1.64	-0.86	1.64
17	-1.29	1.50	-1.29	1.50
18	-0.43	1.48	-0.43	1.48
19	-0.27	0.94	-0.27	0.94
20	-0.91	1.44	-0.91	1.44
21	-0.55	0.85	-0.55	0.85
22	-0.24	0.51	-0.24	0.51
23	-0.85	1.12	-0.85	1.12
24	-0.99	1.15	-0.99	1.15
25	-0.62	0.81	-0.62	0.81
26	-1.19	1.00	-1.19	1.00
27	-0.72	0.97	-0.72	0.97
28	0.28	1.37	0.28	1.37
29	0.00	1.36	0.00	1.36
30	-0.17	1.50	-0.17	1.50
31	-0.22	0.90	-0.22	0.90
32	-0.47	0.44	-0.47	0.44
33	-0.42	0.76	-0.42	0.76
34	1.62	0.31	1.62	0.31
35	-0.89	0.45	-0.89	0.45
36	0.29	1.59	0.29	1.59
37	0.82	1.59	0.82	1.59
38	0.73	0.91	0.73	0.91
39	0.67	1.48	0.67	1.48
40	-0.24	0.75	-0.24	0.75

Resultados

Con la finalidad de evaluar el efecto de purificación del test en la detección del DIF, se han tenido en cuenta tanto el porcentaje de ítems con DIF correctamente identificados (IC), como el porcentaje de ítems que sin presentar DIF han sido detectados como tales, es decir, el porcentaje de falsos positivos (FP). Estos porcentajes han sido calculados a través de las 50 réplicas simuladas. En las tablas 3 a 6 se presentan los resultados obtenidos para cada una de las medidas de DIF utilizadas y en cada una de las condiciones manipuladas. Además, estos resultados se presentan en función del nivel de significación considerado (5% y 1%).

El uso del procedimiento de igualación bietápico, en general e independientemente del estadístico utilizado para evaluar el DIF, no mejoró la tasa de IC, el porcentaje de IC fue del 71.63% en el primer paso frente a un 66% cuando el test fue purificado. En el estadístico de Lord y al nivel de significación del 5%, el porcentaje promedio de IC fue del 77% para el procedimiento no iterativo y del 72% cuando se purificó el test; en la medida de área con signo, el porcentaje medio de IC fue del 58% en el primer paso y del 52% en el segundo; por último, en la medida de área sin signo, el porcentaje de ítems correctamente identificados fue del 80% sin purificar el test frente a un 74% cuando el test fue purificado. Los resultados encontrados están afectados por el tamaño muestral y la cantidad de DIF manipulado. Así, en las condiciones de tamaño muestral alto (N=1000) y mayor cantidad de DIF (d=0.60), el porcentaje de IC fue el mismo en los dos procedimientos (no iterativo y bietápico), siendo además tasas elevadas, los valores encontrados fueron del 100% o cercanos al mismo (ver tabla 3). En las condiciones de menor tamaño muestral y menor cantidad de DIF el uso de un procedimiento de purificación del test no mejoró la tasa de identificaciones correctas.

De los tres estadísticos utilizados para evaluar el DIF, χ^2 de Lord y Z(H) de Raju fueron los que obtuvieron un mayor porcentaje de IC frente a la prueba Z(SA). Sin embargo, estos resultados deben ser considerados teniendo en cuenta el tipo de DIF evaluado (ver tablas 3 y 4). Cuando el DIF manipulado fue uniforme y no uniforme asimétrico, las tasas de IC fueron similares para las tres medidas utilizadas, resultados que no se mantienen, tal y como era de esperar, cuando el DIF fue no uniforme simétrico.

En cuanto al tamaño muestral, e independientemente de si el test fue o no purificado, los resultados encontrados muestran que el número de ítems correctamente identificados aumenta conforme el tamaño muestral es más elevado, produciéndose un efecto techo cuando la muestra fue de 1000 sujetos.

De acuerdo con lo esperado, una mayor diferencia entre los parámetros del ítem del grupo de referencia y los del grupo focal implica una mayor tasa de IC (tablas 3 y 4). Esta tendencia se mantuvo independientemente del estadístico de DIF utilizado, el tamaño muestral, tipo de DIF manipulado y porcentaje de ítems con DIF en el test.

El porcentaje de IC varió en función del nivel de significación fijado. Cuando el tamaño muestral es pequeño y la diferencia en DIF es menor, el porcentaje de ítems correctamente identificados disminuye conforme el nivel de probabilidad es más restrictivo. Sin embargo, en las condiciones de mayor tamaño muestral y cantidad de DIF, ésta pauta no se presenta, del tal modo que los ítems correctamente identificados al 5%, también lo son a niveles de significación más bajos. Si observamos los datos de las tablas 3 y 4, en las condiciones de N=1000 y d=0.6, cuando el DIF manipula-

do fue uniforme y no uniforme asimétrico, el porcentaje de IC es del 100%, tanto al 5% como al 1%, aunque cuando el DIF es no uniforme simétrico las tasas de IC fueron algo menores al 1% que al 5%, pero en cualquiera de los casos estas fueron del 90% o superiores, tanto para el estadístico de Lord como para Z(H).

Aunque, los resultados no mostraron una mejora del porcentaje de IC del procedimiento bietápico de igualación frente al procedimiento no iterativo de igualación, si se observó (tablas 5 y 6) una disminución del porcentaje de falsas identificaciones. En promedio, el porcentaje de FP, al nivel de significación del 5%, fue del 12.10% sin purificar el test frente a un 6.87% cuando el test se purificó. A un nivel de significación más restrictivo (1%), los resultados fueron 3.91% de FP en la primera etapa frente a un 1.22% en la segunda etapa. Esta tendencia se mantuvo independientemente del estadístico utilizado para detectar el DIF. De este modo, al nivel de significación del 5%, para el estadístico de Lord, el porcentaje promedio de FP fue del 10.41% en la primera etapa disminuyendo hasta un 5.44%, para Z(SA) la reducción de FP fue del 11.50% al 6.05%, y para Z(H) fue del 14.39% al 9.11%. La pauta anterior se mantiene si consideramos el nivel de significación del 1%.

El número de detecciones incorrectas estuvo afectado por el porcentaje de ítems con DIF en el test, siendo el efecto de la purificación del test más relevante cuando el test contiene mayor número de ítems con DIF. En general, al nivel de significación del 5%, el porcentaje promedio de FP fue del 5.05% (χ^2 de Lord), 6.47% (Z(SA)) y 9% (Z(H)) cuando el test contenía un 10% de los ítems con DIF, y del 10.35% (χ^2 de Lord), 11.07% (Z(SA)) y 14.5% (Z(H)) cuando el test contenía un 30%. Es más, considerando el mismo nivel de significación, el porcentaje promedio de FP en la primera etapa del proceso de purificación y cuando el test contenía un 10% de los ítems con DIF, fue de 6.33% (χ^2 de Lord), 7.49% (Z(SA)) y 10.16% (Z(H)) disminuyendo, respectivamente, a los niveles del 4.68%, 5.45% y 7.84%. Cuando el test contenía un 30% de los ítems con DIF, los porcentajes promedios de FP en el primer paso fueron 14.89% (χ^2 de Lord), 15.5% (Z(SA)) y 18.63% (Z(H)), y en el segundo fueron 6.20%, 6.43% y 10.37%. Al nivel de significación del 1%, el porcentaje de promedio de FP en tests con 10% de ítems con DIF, fue de 1.32% (χ^2 de Lord), 1.68% (Z(SA)) y 2.06% (Z(H)) cuando no se purificó el test y del 0.71%, 0.91% y 0.95% cuando se purificó el test. En tests con el 30% de los ítems con DIF, el porcentaje promedio en el primer paso del proceso fue de 5.75% para el estadístico de Lord, 6.21% para Z(SA) y 6.45% para Z(H); en el segundo paso del proceso fueron 1.42%, 1.43% y 1.92%, respectivamente para cada uno de los estadísticos utilizados.

El estadístico de Lord y Z(SA) frente a Z(H) controlaron mejor la tasa de FP. Además, el número de falsos positivos, tal y como era de esperar, también aumentó al incrementar el tamaño muestral. Por último, la tasa de FP disminuyó en todos los tamaños

Tabla 3
Porcentajes de identificaciones correctas al nivel de significación del 5%

Tipo de DIF	Tamaño muestral	Cantidad de DIF	% ítems con DIF	Lord		Z(SA)		Z(H)	
				1ª	2ª	1ª	2ª	1ª	2ª
Uniforme	250	0.4	10%	34	31	42	36	40	36
			30%	51.5	36	50.5	42	57	42.5
		0.6	10%	72	64	74	66	73	72
			30%	89.5	67.5	85.5	68.5	90	70.5
	500	0.4	10%	52	48	58	50	59	52
			30%	76	55	75.5	53.5	81	60.5
		0.6	10%	93	88	85	82	91	90
			30%	98.5	87.5	98	82	99	88.5
	1000	0.4	10%	95	92	88	83	93	89
			30%	98	85.5	95	81	98.5	85
		0.6	10%	100	100	99	98	100	100
			30%	100	100	100	99	100	99.5
No Uniforme	250	0.4	10%	18	18	0	0	24	22
			30%	37	27	9	5	30.5	23.5
		0.6	10%	34	36	8	10	40	40
			30%	65.5	56	15	5	61	49
	500	0.4	10%	56	56	14	10	64	62
			30%	68	55	6	3.5	65	48.5
		0.6	10%	82	80	4	2	90	82
			30%	92.5	92	11	5	92	89
	1000	0.4	10%	80	84	10	8	56	80
			30%	92.5	90.5	16	8.5	94	85
		0.6	10%	100	100	6	6	100	100
			30%	98.5	98.5	22	7.5	99.5	98.5
Mixto	250	0.4	10%	26	20	42	34	40	34
			30%	58.5	53	65	62.5	84	63.5
		0.6	10%	68	60	70	66	80	72
			30%	96.5	87	97	88.5	98	91
	500	0.4	10%	62	54	60	58	76	66
			30%	95	80.5	95.5	82.5	97	87.5
		0.6	10%	98	94	96	88	100	98
			30%	100	98.5	100	100	100	99.5
	1000	0.4	10%	92	90	92	88	96	94
			30%	100	95.5	99.5	98.5	100	98
		0.6	10%	100	100	100	100	100	100
			30%	100	100	100	100	100	100

muestrales y condiciones manipuladas con el incremento del nivel de confianza (ver tabla 6).

Discusión

Los resultados de este trabajo muestran que el procedimiento bietápico de igualación produce una disminución del porcentaje de falsos positivos, acercándolos a los niveles nominales, aunque no mejora la identificación correcta de ítems con DIF. Las conclusiones de este estudio son consistentes con las de otros, donde se muestra que la utilización de un procedimiento de purificación del test mejora la tasa de falsos positivos, pero no necesariamente nos lleva a mejorar la detección correcta de ítems con DIF (Kim y Cohen, 1992; Miller y Oshima, 1992; Lautenschlager, Flaherty y Park, 1994).

Tabla 4
Porcentajes de identificaciones correctas al nivel de significación del 1%

Tipo de	Tamaño	Cantidad	% ítems	Lord		Z(SA)		Z(H)	
DIF	muestral	de DIF	con DIF	1ª	2ª	1ª	2ª	1ª	2ª
Uniforme	250	0.4	10%	16	13	18	14	15	13
			30%	29	13.5	31.5	17.5	32	14
		0.6	10%	53	41	51	44	59	43
			30%	69	42.5	70	45	71.5	42
	500	0.4	10%	30	24	26	21	29	22
			30%	56.5	34.5	50.5	28.5	56	33.5
		0.6	10%	82	77	77	72	78	76
			30%	93	69.5	88	64	92	65.5
	1000	0.4	10%	78	70	76	69	74	69
			30%	90.5	66.5	87	63	89.5	62
		0.6	10%	100	99	97	92	100	99
			30%	100	98.5	99.5	96	100	97.5
No uniforme	250	0.4	10%	4	4	0	0	2	2
			30%	16	12	1.5	0.5	5	2
		0.6	10%	14	16	0	0	8	8
			30%	38.5	20.5	6	0.5	23.5	7
	500	0.4	10%	26	26	2	0	22	24
			30%	46	32.5	0.5	1	27.5	18
		0.6	10%	54	60	2	0	46	58
			30%	83	71	4	1.5	68.5	57.5
	1000	0.4	10%	56	60	4	2	42	44
			30%	86.5	77	7	2.5	78	59.5
		0.6	10%	90	94	2	0	84	92
			30%	97	96	8.5	1.5	98	95
Mixto	250	0.4	10%	4	4	16	4	8	4
			30%	48.5	32	51.5	35.5	54.5	34
		0.6	10%	36	32	54	38	36	32
			30%	90	75	88.5	78.5	91.5	77
	500	0.4	10%	38	36	38	34	40	36
			30%	84.5	59	81	63.5	86.5	62.5
		0.6	10%	90	88	78	72	94	90
			30%	100	95.5	100	96.5	100	96
	1000	0.4	10%	76	72	82	68	82	76
			30%	99.5	89.5	99	93.5	100	90.5
		0.6	10%	100	100	100	100	100	100
			30%	100	100	100	100	100	100

Tabla 5
Porcentajes de falsos positivos al nivel de significación del 5%

Tamaño	**Cantidad**	**% ítems**	**Lord**		**Z(SA)**		**Z(H)**	
Muestral	**de DIF**	**con DIF**	**1ª**	**2ª**	**1ª**	**2ª**	**1ª**	**2ª**
250	0.4	10%	4.5	3.67	7.39	4.83	9	8
		30%	8.36	3.57	9.29	4.5	12	7.79
	0.6	10%	6.78	3.89	7.78	5	9.39	6.56
		30%	13.57	4.79	15.93	7.43	17.21	8.86
500	0.4	10%	5.61	4.22	6.56	5.06	8.72	7.11
		30%	8.5	6.07	8.5	5.36	12.5	9.64
	0.6	10%	6.17	5.72	7.89	5.72	10.78	8.33
		30%	13.07	6.64	15.57	8	18	11
1000	0.4	10%	6.83	4.78	7.78	5.5	10.67	8
		30%	16.64	7.57	16.64	5.86	21.93	11.5
	0.6	10%	8.11	5.78	7.56	6.61	12.39	9.06
		30%	26.79	8.57	27.07	8.71	30.14	13.43

Tabla 6
Porcentajes de falsos positivos al nivel de significación del 1%

Tamaño Muestral	Cantidad de DIF	% ítems con DIF	Lord		Z(SA)		Z(H)	
			1ª	2ª	1ª	2ª	1ª	2ª
250	0.4	10%	0.89	0.39	1.22	0.72	1.5	0.78
		30%	1.64	0.79	2.64	0.79	2.64	0.64
	0.6	10%	1.11	0.56	1.89	0.61	2	0.5
		30%	5.43	0.79	5.79	1.21	6	1.29
500	0.4	10%	1.22	0.89	1.17	0.72	1.44	0.83
		30%	2.14	1.07	2.5	1.00	3.21	1.79
	0.6	10%	1.61	0.83	1.72	1.17	2.33	1.28
		30%	5.14	1.57	6.71	2.07	7.07	2.29
1000	0.4	10%	1.28	0.78	1.89	1.11	2.44	1.17
		30%	6.14	1.71	6.21	0.93	6.07	2.14
	0.6	10%	1.78	0.83	2.17	1.11	2.67	1.11
		30%	14.0	2.57	13.43	2.57	13.71	3.36

El procedimiento bietápico de igualación mostró mejores resultados cuando el porcentaje de ítems con DIF en el test fue mayor (30%). En esta situación los porcentajes de FP, cuando no se utiliza ninguna purificación del test, fueron superiores a los niveles nominales esperados, pero éstos descendieron cuando se aplicaron las tres medidas de forma iterativa.

En cuanto a la eficiencia de las medidas utilizadas para evaluar el DIF, el estadístico de Lord y Z(H) fueron más eficaces en la identificación del DIF que Z(SA), resultados que concuerdan con los encontrados por Cohen y Kim (1993) e Hidalgo y López (1997a). Si consideramos la detección de ítems con DIF uniforme y no uniforme asimétrico, las tasas de identificaciones correctas fueron similares en las tres medidas; sin embargo, la Z(SA) es poco eficaz para detectar DIF no uniforme simétrico (Cohen y Kim, 1993; Raju, 1990). El estadístico de Lord presentó un mejor control de las tasas de FP, siendo sistemáticamente menores a las obtenidas cuando se aplicó Z(SA) y Z(H). Aunque los resultados de este estudio reflejan la ventaja de utilizar el estadístico de Lord frente a las medidas de Raju, lo recomendable es utilizar conjuntamente ambas aproximaciones, dado que proporcionan información complementaria en la evaluación del DIF (Cohen y Kim, 1993).

En general, de acuerdo con lo esperado, tanto las tasas de identificaciones correctas como las de falsos positivos aumentaron conforme aumenta el tamaño muestral, la cantidad de DIF y el número de ítems con DIF en el test. A la vista de los resultados, se aconseja la utilización de niveles de significación del 1% , dado que el porcentaje de FP se reduce sin disminución del porcentaje de IC, siempre y cuando trabajemos con tamaños muestrales grandes (N=1000).

Aunque las conclusiones generales de este trabajo apoyan la utilización del procedimiento bietápico de igualación, es necesario obtener mayor evidencia acerca de su comportamiento en situaciones de impacto entre los grupos a comparar, tamaños muestrales no equivalentes entre el grupo focal y el de referencia, tamaños muestrales pequeños y tests con mayor o menor porcentaje de ítems con DIF que los contemplados en este estudio.

Referencias

Baker, F.B. (1993). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement, 17,* 20.

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

Cohen, A.S. y Kim, S.H. (1993). A comparison of Lord's χ^2 and Raju's area measures in detection of DIF. *Applied Psychological Measurement, 17,* 39-52.

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., Mellenbergh, G.J. y Muñiz, J. (1998). Comparación del procedimiento Mantel-Haenszel frente a los modelos loglineales en la detección del funcionamiento diferencial de los ítems. *Psicothema, 10,* 209-218.

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

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,* 397-411.

Hidalgo, M.D. y López, J.A. (1995). *SIMULA 2.0: Un programa para la simulación de vectores de respuesta al ítem.* Demostración de software presentada al IV Symposium de Metodología de las CC. del Comportamiento, La Manga, Murcia.

Hidalgo, M.D. y López, J.A. (1997a). 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.

Hidalgo, M.D. y López, J.A. (1997b). *Detección del DIF en ítems politómicos e igualación iterativa: comparación entre las medidas de área de Raju y el estadístico de Lord.* Comunicación presentada en el V Congreso de Metodología de las CC. Humanas y Sociales, Sevilla.

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

Kim, S.H. y Cohen, A.S. (1992). IRTDIF: A computer program for IRT differential item functioning analysis. *Applied Psychological Measurement, 16*, 158.

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

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

Lautenschlager, G.J. y Park, D. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness and parameter linking. *Applied Psychological Measurement, 12,* 365-376.

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

Mellenbergh, G.J. (1982). Contigency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

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

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

Mislevy, R.J. y Bock, R.D. (1990). *PC-BILOG 3: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software.

Park, D. y Lautenschlager, G.J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement, 14,* 163-173.

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

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

Stocking, M.L. y Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

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

Aceptado el 25 octubre de 1999