Psicothema, 1997. Vol. 9, nº 2, pp. 417-431
ISSN 0214 - 9915 CODEN PSOTEG

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

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

El presente estudio pretende identificar las condiciones bajo las cuales la medida de área con signo Z(AES), la medida de área sin signo Z(AEA), el estadístico de Lord y el análisis de regresión logística (RL) detectan la presencia de funcionamiento diferencial del ítem (FDI). Las condiciones manipuladas fueron: tamaño muestral, cantidad y tipo de FDI, presencia o no de impacto, porcentaje de items con FDI en el test y nivel de significación. Los resultados encontrados muestran que el estadístico de Lord y Z(AEA) son bastante eficaces en la detección correcta de FDI uniforme, no uniforme y mixto. Por otro lado, el procedimiento de RL fue eficaz en la detección del FDI no uniforme y mixto, pero no en la del FDI uniforme. El estadístico de Lord, Z(AES) y Z(AEA) obtuvieron tasas de falsos positivos más elevadas que RL, principalmente cuando el tamaño muestral y el porcentaje de items con FDI en el test fueron elevados. Esto también ocurrió cuando la cantidad de FDI fue alta.

An comparison of the Area Methods, Lord's chi-square test and Logistic Regression Analysis for Assesing Differential Item Functioning. The present study compared the performance under different conditions of the signed area measure Z(AES), the unsigned area measure Z(AEA), Lord's chi-square test and the logistic regression analysis (LR) on detection of differential item functioning (DIF). Sample size, amount and type DIF, difference in the group trait level averages, percent of DIF items test and significance levels were manipulated. The results show that Lord's test and Z(AEA) were effective in assessing uniform, non-uniform and mixed DIF. The LR procedure was low power for detecting uniform DIF, however it was able to detect non-uniform and mixed DIF. When the sample size, amount of DIF and percent of DIF items were larger high number of false positives were obtained using the Lord's statistic, Z(AES) and Z(AEA).

Una de las críticas más importantes a los tests psicométricos ha sido afirmar que están sesgados, es decir, que favorecen injustamente a los sujetos de un grupo (p.e. clase social alta, varones...) sobre los sujetos de otro grupo (p.e. clase social baja, mujeres...) identificando erróneamente diferencias entre grupos. Cuando se utiliza un test en la evaluación de cualquier rasgo o habilidad

Correspondencia: Mª Dolores Hidalgo Montesinos
Facultad de Psicología
Campus de Espinardo. Apdo. 4.071
Universidad de Murcia
30080 Murcia (Spain)

psicológica es de esperar que éste sirva al propósito por el cual fue construido y proporcione medidas fiables y válidas acerca de la manifestación del rasgo en los sujetos. Así, si se administra un test a dos grupos de sujetos que difieren en el rasgo evaluado, estas diferencias (impacto) deben ser detectadas por el test. Del mismo modo, si los grupos no difieren (no impacto) las puntuaciones obtenidas en el test deben reflejar la igualdad entre grupos. Esto se traduce a que las propiedades psicométricas del test y por ende de los items que lo conforman sean invariantes a través de distintos grupos o muestras de una misma población. En ocasiones, y contrario a lo esperado, algunos de los items de un test pueden funcionar diferencialmente en los grupos en los que se han administrado, siendo necesario estudiar este hecho, que en el ámbito psicométrico se ha denominado Funcionamiento Diferencial del Item (FDI). El FDI se produce cuando, en dos o más grupos equivalentes, la probabilidad de obtener una respuesta correcta, dado un nivel de habilidad es diferente para cada uno de dichos grupos (Scheuneman, 1979).

Normalmente, en los estudios de FDI, los sujetos son clasificados en dos grupos: Focal (F) y Referencia (R). Se denomina grupo focal al grupo objeto de análisis, casi siempre un grupo minoritario. Por contra, el grupo de referencia se toma como grupo base o de comparación, casi siempre un grupo mayoritario. La variable de agrupamiento podría ser cualquiera sociodemográfica (sexo, etnia, edad, nivel educativo...) en la que se sospeche que las propiedades psicométricas de los items pueden diferir.

Mellenbergh (1982) definió dos tipos de FDI: Uniforme cuando la probabilidad de responder correctamente a un ítem es mayor en un grupo que en otro y No uniforme cuando esta probabilidad es mayor en un grupo que en otro, hasta un nivel de habilidad dado, y a partir de dicho nivel de habilidad las probabilidades se invierten siendo menores en el primer grupo que en el segundo. En este último caso se dan dos situaciones: No uniforme propiamente dicho cuando las diferencias en probabilidad entre los grupos sometidos a análisis se cancelan, y Mixto cuando estas diferencias no se anulan.

Las técnicas propuestas para evaluar el FDI se pueden clasificar como (Millsap y Everson, 1993): a) Métodos de Invarianza Condicional Observada (ICO) que utilizan las puntuaciones observadas en el test como variable de equiparación. Aquí se podría incluir entre otros el estadístico de Mantel-Haenszel (Holland y Thayer, 1988), los modelos logit (Mellenbergh, 1982) y el análisis de regresión logística (Rogers y Swaminathan, 1993; Swaminathan y Rogers, 1990); y b) Métodos de Invarianza Condicional No observada (ICN), donde se trabaja a partir de las puntuaciones de habilidad estimadas según algún modelo de medida. En la Teoría de la Respuesta a los Items se han propuesto métodos de comparación de parámetros (Lord, 1980), medidas de área (Cohen, Kim y Baker, 1993; Kim y Cohen, 1991; Raju, 1988, 1990) y métodos basados en la comparación de modelos (Thissen, Steinberg y Wainer, 1988, 1993).

Ante tal cantidad de aproximaciones, puede resultar útil conocer qué procedimientos son los más eficaces en la evaluación del FDI y bajo qué condiciones. Desde los métodos de ICO, el estadístico de Mantel-Haenszel (MH), por su sencillez de cálculo e interpretación, ha sido uno de los procedimientos más utilizados e investigados. Sin embargo, aunque es eficaz en la detección de FDI uniforme, es incapaz de detectar correctamente el FDI no uniforme, salvo cuando se aplican variaciones iterativas del mismo (Clauser, Mazor y Hambleton, 1993; Fidalgo, 1996). El análisis de regresión logística (RL) más complejo y costoso de realizar, es más potente en la identificación co-

rrecta de FDI no uniforme que el estadístico de MH. Las ventajas del análisis de RL no sólo radican en lo anteriormente comentado, sino que frente al procedimiento de MH, RL establece una relación funcional entre la respuesta al ítem y la variable de comparación. Una característica de los métodos ICO es que no establecen ningún supuesto sobre el modelo de medida subyacente a los datos del test. Estos métodos son aplicados tomando las puntuaciones observadas en el test. Por contra, algunos de los métodos de ICN se aplican una vez ajustado un modelo de TRI. Desde un punto de vista teórico resulta más apropiado trabajar bajo este tipo de modelos dadas las ventajas estadísticas que se derivan del ajuste de los mismos (Lord, 1980). Sin embargo, en la práctica no siempre es posible implementar estos modelos dado que el buen ajuste de los mismos requiere, entre otras cosas, grandes muestras de sujetos que rara vez se dispone en la investigación aplicada sobre FDI. Es más, estos métodos son menos fáciles de aplicar que MH o RL.

Una parte de los trabajos sobre FDI se han ocupado de estudiar el acuerdo entre los métodos de ICO y de ICN (Hambleton y Rogers, 1989; Hidalgo, 1995; Navas y Gómez, 1994; Raju, 1990). Así, Hambleton y Rogers (1989) encontraron que el estadístico de MH y las medidas de área proporcionan resultados más o menos similares en la detección correcta de FDI uniforme. Mientras que, cuando el FDI fue no uniforme las medidas de área identificaron correctamente mayor cantidad de items con FDI que el estadístico de MH. Raju (1990) encontró que las medidas exactas de área sin signo se mostraron más precisas en la evaluación del FDI frente a las medidas de área con signo y el estadístico de MH. De los trabajos de Hambleton y Rogers (1989) y Raju (1990) se deduce la preferencia de los procedimientos basados en TRI frente a métodos de ICO tales como MH. Sin embargo, pocos son los trabajos que han comparado RL (más potente que MH en la detección del FDI no uniforme) y los procedimientos de TRI (Hidalgo, 1995). Navas y Gómez (1994) compararon entre otras técnicas RL y las medidas exactas de área de Raju en la detección de FDI uniforme, encontrando que ambos procedimientos detectan por igual la presencia de FDI. No se conoce el acuerdo entre estas técnicas en la evaluación del FDI no uniforme y tampoco ha sido estudiado su comportamiento en distintas condiciones de porcentaje de items con FDI en el test, cantidad de FDI, y tipo de FDI.

El presente trabajo pretende identificar las condiciones (tamaño muestral, porcentaje de items con FDI en el test, cantidad y tipo de FDI) bajo las cuales el análisis de RL, las medidas exactas de área de Raju (1990) y el estadístico de Lord (1980) detectan mejor la presencia de FDI. Para ello se realizó un estudio de simulación.

Regresión Logística

La ecuación general para un modelo de RL vendría dada por (Hosmer y Lemeshow, 1989):

$$p(y=1|x)=\frac{e^{f(x)}}{1+e^{f(x)}} \qquad (1)$$

donde y es la variable de respuesta, p(y=1|x) es la probabilidad de obtener una respuesta correcta (probabilidad de éxito) condicionado a x, x es el vector de variables predictoras y $f(x)= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$, siendo p el número de variables predictoras. El modelo no lineal de la ecuación 1 puede transformarse a un modelo lineal aditivo, efectuando sobre la variable criterio una transformación logit. En el estudio del FDI, $f(x)= \beta_0 + \beta_1 H + \beta_2 G + \beta_3 HG$. Donde β_0 es el efecto total de la dificultad del ítem, β_1 es el efecto de la variable H, que se define co-

mo la puntuación observada en el test, β_2 es el efecto de la variable grupo (G) y β_3 es el efecto de la interacción habilidad x grupo (HG). Un ítem mostrará FDI uniforme si $\beta_2 \neq 0$ y $\beta_3=0$, y FDI no uniforme si $\beta_3 \neq 0$, independientemente que β_2 sea igual a cero o no (Swaminathan y Rogers, 1990). Estas hipótesis, normalmente, se prueban utilizando el estadístico de Wald que permite probar la significación de los pesos β comparando el valor estimado ($\hat{\beta}_p$) para un peso p con su error típico (SE ($\hat{\beta}_p$)) según la siguiente expresión:

$$W = \left(\frac{\hat{\beta}_p}{SE(\hat{\beta}_p)} \right)^2 \qquad (2)$$

que sigue una distribución χ_1^2.

Medidas de Raju

En la TRI se han propuesto un conjunto de medidas basadas en el cálculo del área entre la Curva Característica del Item (CCI) en el grupo focal y la CCI en el grupo de referencia. Raju (1988, 1990) propone dos medidas de área exactas (con signo y sin signo) disponibles en los modelos de 1-p, 2-p y 3-p que permiten probar su significación estadística a través de una prueba Z. La expresión general de estas medidas, basadas en la integración continua, viene dada por:

$$A_i = \int_{\infty}^{\infty} f[P_R(\theta) - P_F(\theta)] \; d(\theta) \qquad (3)$$

donde $P_R(\theta)$ y $P_F(\theta)$ representan las probabilidades de responder correctamente al ítem *i* en cada uno de los grupos (R y F). La función *f* puede especificarse con signo o sin signo. En el primer caso se obtendría la medida de área exacta con signo (AES), y en el segundo una medida de área exacta absoluta (AEA). La expresión general de la ecuación 3 adopta distintas formas según el modelo de TRI con el que se este trabajando, y si se cumplen o no ciertas condiciones en los parámetros de los items. Así, en el modelo de 2-p, Raju (1988, 1990) define el área con signo como AES= ($\hat{b}_{iF} - \hat{b}_{iR}$) y el área absoluta como AEA= $|\hat{b}_F - \hat{b}_R|$ cuando los parámetros de discriminación son iguales en los grupos focal y de referencia o AEA= $|H_i|$ cuando son distintos. El valor H(i) se obtiene según:

$$H_i = \frac{2(\hat{a}_{iF} - \hat{a}_{iR})}{D\hat{a}_{iR}\hat{a}_{iF}} \; 1n\left(1 + \exp\left(\frac{D\hat{a}_{iR}\hat{a}_{iF}(\hat{b}_{iF} - \hat{b}_{iR})}{\hat{a}_{iF} - \hat{a}_{iR}}\right)\right) - (\hat{b}_{iF} - \hat{b}_{iR}) \qquad (4)$$

Las pruebas estadísticas para las medidas de Raju (1988) son:

Prueba de significación para AES. Se asume que AEA se distribuye normalmente. De este modo, para probar estadísticamente si las diferencias entre dos CCIs son significativas se puede utilizar la siguiente prueba Z (Raju, 1990, p. 202):

$$Z_i(AES) = \frac{AES}{\left(\sigma^2_{b_{iF}} + \sigma^2_{b_{iR}}\right)^{1/2}} \qquad (5)$$

donde $\sigma^2_{b_{iF}}$ y $\sigma^2_{b_{iR}}$ vienen dadas en Raju (1990, pp.200).

Prueba de significación para AEA. No es posible establecer el supuesto de normalidad para las medidas AEA. Cuando $\hat{a}_{iR} \neq \hat{a}_{iF}$ Raju (1990, pp. 203) recomienda que la prueba de significación se realice sobre H y no sobre su valor absoluto. Así,

$$Z_i(AEA) = \frac{H_i}{\sigma_{Hi}} \qquad (6)$$

Cuando $\hat{a}_{iR} = \hat{a}_{iF}$ se utiliza la prueba estadística de la ecuación 5.

A fin de probar la significación estadística de estas medidas exactas de área, el valor Z se compara con el valor teórico correspondiente a la distribución normal tipificada, dado el nivel de confianza prefijado por el investigador. Si el valor Z observado es mayor o igual que el valor teórico, el ítem bajo estudio presenta FDI. En caso contrario, se considera que el ítem no presenta FDI.

Procedimiento de Lord

Un ítem funciona diferencialmente en dos grupos si los parámetros que lo definen varían a través de los grupos. Lord (1980) propone un estadístico que somete a comprobación esta hipótesis. Este estadístico viene dado por (Lord, 1980, p. 233):

$$\text{LORD} - \chi^2 = V'S^{-1}V \qquad (7)$$

donde V es el vector de diferencias entre los parámetros estimados para un ítem en el grupo de referencia y los parámetros estimados para ese mismo ítem en el grupo focal. Bajo el modelo de 2-p, V' es:

$$V' = (\hat{b}_R - \hat{b}_F, \hat{a}_R - \hat{a}_F) \qquad (8)$$

y S^{-1}, es la inversa de la matriz de varianza-covarianza asintótica para los vectores de diferencias entre parámetros.

El estadístico propuesto por Lord, bajo la hipótesis nula, sigue una distribución χ^2 con dos grados de libertad. Un ítem presenta FDI si el valor observado LORD – χ^2 es mayor que el valor teórico asociado a χ^2_2 al nivel de significación establecido.

Método

Condiciones experimentales

Se han seleccionado tres tamaños muestrales de 250, 500 y 1000 sujetos tanto para el grupo focal como para el grupo de referencia y un tamaño de test fijo (75 items). Para cada uno de los tamaños muestrales se generaron dos distribuciones de habilidad normales en el intervalo [-3, +3] con igual varianza ($\sigma^2_\theta = 1$) y distinta media (caso 1: $\mu_\theta = 0$ y caso 2: $\mu_\theta = -1$). Esto proporcionó dos situaciones: no impacto donde las medias de ambos grupos (F y R) no difieren, e impacto donde la media del grupo focal fue de -1.

Para el conjunto de 75 items que forman el test bajo estudio se generaron aleatoriamente valores de dificultad y de discriminación. Los valores de discriminación se simularon para que adoptaran una distribución uniforme entre los límites [0.3, 2] y los de dificultad a partir de una distribución normal N(0,1) cuyos límites varían entre [-2.3, 2.3].

Para cada uno de los tamaños muestrales y situaciones (no impacto e impacto) se establecieron 9 condiciones donde se manipuló la cantidad de items con FDI en el test, el tipo de FDI (uniforme, no uniforme y mixto) y la cantidad de FDI (definido como diferencia, d_{R-F}, entre los parámetros de dificultad y/o discriminación de los grupos a comparar). Las condiciones manipuladas fueron: 1) 20% de los items con FDI y d_{R-F}= 0.4, 2) 20% de los items con FDI y d_{R-F}= 0.7, 3) 20% de los items con FDI y d_{R-F}= 1.0, 4) 33% de los items con FDI y d_{R-F}= 0.4, 5) 33% de los items con FDI y d_{R-F}= 0.7, 6) 33% de los items con FDI y d_{R-F}= 1.0, 7) 40% de los items con FDI y d_{R-F}= 0.4, 8) 40% de los items con FDI y d_{R-F}= 0.7 y 9) 40% de los items con FDI y d_{R-F}= 1.0.

En cada una de estas condiciones el tipo de FDI generado fue en el mismo sentido. En todas las condiciones el número de items con FDI uniforme, no uniforme y mixto fue el mismo.

Generación de las matrices de datos

Según lo expuesto en el apartado anterior y teniendo en cuenta los 3 tamaños mues-

trales, los 2 tipos de distribución de habilidad y las 9 condiciones se dispone de un total de 54 combinaciones posibles con respecto al grupo focal. A cada una de estas condiciones corresponde una matriz de datos (sujetos x items). Esta matriz fue generada con el programa SIMULA v. 2 (Hidalgo y López, 1995) bajo el modelo logístico de 2-p.

Con la finalidad de encontrar resultados estables en cada una de las 54 combinaciones posibles, se obtuvieron 10 réplicas, sometiendo a estudio un total de 540 matrices. Para establecer la comparación correspondiente entre grupo de referencia y grupo focal fueron generadas 10 réplicas más por cada tamaño muestral a partir de la distribución normal N(0,1) de habilidad y de los valores iniciales de los parámetros de los items.

Detección del FDI

En el análisis de RL el FDI se evaluó en cada ítem mediante el modulo complementario LOGIT del paquete SYSTAT (Steinberg y Phillips, 1991). Las variables independientes definidas fueron: la puntuación observada del sujeto en el test tratada como un predictor continuo y la pertenencia a grupo. La variable dependiente fue la respuesta al ítem de naturaleza dicotómica.

Tanto en la aplicación del estadístico de Lord como de las medidas de área de Raju se estimaron, primeramente, los parámetros de los items en los grupos focal y de referencia, separadamente. Las estimaciones fueron realizadas con el programa BILOG versión 3.04 (Mislevy y Bock, 1990) utilizando las opciones por defecto del mismo. A continuación los parámetros estimados en ambos grupos fueron igualados utilizando el programa EQUATE versión 2.0 (Baker, 1993) que implementa el procedimiento de curvas características desarrollado por Stocking y Lord (1983). Por último, se calculó LORD – χ^2, Z(AES) y Z(AEA) con el programa IRTDIF (Kim y Cohen, 1992b) que permite obtener las medidas anteriores.

Resultados

A fin de evaluar la eficacia de los procedimientos empleados, se han tenido en cuenta tanto el porcentaje de items con FDI correctamente identificados (IC), como el porcentaje de items que sin presentar FDI han sido detectados como tales, es decir, el porcentaje de falsos positivos (FP) a través de las 10 réplicas analizadas. En las tablas 1 a la 8 se presentan los resultados obtenidos en cada una de las condiciones manipuladas y estadísticos de FDI calculados. Estos aparecen en tres niveles de significación: 5%, 1% y 0.1% y resumidos en función del tipo de FDI generado: uniforme, no uniforme y mixto.

Situación de no impacto

Regresión Logística. Se puede observar que conforme aumenta la cantidad de FDI generado también aumenta el número de IC independientemente del tipo de FDI, del porcentaje de items con FDI y del tamaño muestral de los grupos focal y de referencia (ver tablas 1 a la 4). Con relación al tipo de FDI generado, el procedimiento de RL mostró, en todas las condiciones, mayor potencia en la detección de FDI no uniforme (diferencias sólo en el parámetro de discriminación) y mixto (diferencias en el parámetro de dificultad y discriminación). La tasa de IC más baja apareció en el caso en que el FDI manipulado era uniforme, aunque ésta se incrementó con el aumento del tamaño muestral y la cantidad de FDI. En las condiciones de mayor tamaño muestral y mayor cantidad de FDI las tasas de IC se situaron en valores similares para ambos tipos de FDI no uniforme, aunque los resultados globales indican que el análisis de RL tiene mayor potencia para detectar FDI no uniforme sobre FDI uniforme o mixto.

Tabla 1
Porcentaje de IC en la situación de no impacto (α= 0.05)

% FDI	FDI	Indice	N= 250			N= 500			N= 1.000		
			d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1
20	U	RL	18	20	38	20	42	58	26	60	68
		χ^2	42	96	98	78	100	98	98	100	100
		Z(AES)	44	82	84	72	92	96	92	100	100
		Z(AEA)	54	86	88	76	94	96	98	100	100
	NU	RL	58	86	96	86	96	100	100	100	100
		χ^2	68	82	92	86	98	98	100	100	100
		Z(AES)	0	2	4	12	6	14	4	80	14
		Z(AEA)	68	90	100	84	96	100	100	100	100
	M	RL	40	78	88	76	96	100	88	100	100
		χ^2	56	90	100	74	100	100	92	100	100
		Z(AES)	34	90	98	54	88	98	84	100	100
		Z(AEA)	62	94	100	86	100	100	96	100	100
33	U	RL	10.00	13.75	26.25	13.75	25.00	47.50	20.00	48.75	70
		χ^2	37.50	85.00	88.75	65.00	93.75	97.50	91.25	100	100
		Z(AES)	36.25	73.75	78.75	58.75	86.25	91.25	78.75	100	100
		Z(AEA)	41.25	77.50	78.75	63.75	87.50	92.50	88.75	100	100
	NU	RL	40.00	73.75	92.50	76.25	95.00	91.25	93.75	100	100
		χ^2	42.50	80.00	83.75	77.50	90.00	92.50	93.75	100	100
		Z(AES)	6.25	10.00	12.50	5.00	16.25	28.75	11.25	30.00	42.50
		Z(AEA)	42.50	83.75	97.50	66.25	91.25	96.25	95.00	100	100
	M	RL	36.25	75.00	82.50	66.25	92.50	96.25	85.00	96.25	98.75
		χ^2	57.50	92.50	96.25	75.00	100	100	87.50	100	100
		Z(AES)	31.25	77.50	96.25	50.00	90.00	95.00	76.25	100	100
		Z(AEA)	58.75	95.00	100	83.75	100	100	91.25	100	100
40	U	RL	11	20	23	18	28	45	11	45	67
		χ^2	41	78	90	67	92	97	81	99	100
		Z(AES)	43	73	80	65	88	91	77	99	100
		Z(AEA)	48	79	84	69	88	92	80	99	100
	NU	RL	46	75	89	75	92	96	92	100	100
		χ^2	49	81	89	70	89	94	94	99	100
		Z(AES)	9	23	32	14	25	54	26	50	67
		Z(AEA)	52	86	95	73	92	99	99	100	100
	M	RL	34	67	79	60	93	96	88	96	100
		χ^2	57	89	99	76	100	100	94	100	100
		Z(AES)	29	75	86	53	86	94	70	87	100
		Z(AEA)	65	96	99	84	100	100	97	100	100

Tabla 2
Porcentaje de IC en la situación de no impacto (α= 0.01)

% FDI	FDI	Indice	N= 250			N= 500			N= 1.000		
			d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1
20	U	RL	2	0	10	2	26	34	12	44	54
		χ^2	32	80	88	64	92	92	90	100	100
		Z(AES)	26	70	72	56	78	88	86	100	100
		Z(AEA)	30	74	72	60	80	88	90	100	100
	NU	RL	40	64	76	78	92	98	100	100	100
		χ^2	42	60	76	72	94	96	96	100	100
		Z(AES)	0	0	0	2	0	6	2	0	4
		Z(AEA)	36	58	94	72	92	100	98	100	100
	M	RL	16	52	70	52	96	100	74	100	100
		χ^2	50	84	100	78	100	100	80	100	100
		Z(AES)	20	66	94	44	82	94	54	100	100
		Z(AEA)	46	88	100	66	100	100	82	100	100
33	U	RL	0	25	10	30	5	13.75	8.75	27.50	48.75
		χ^2	25	67.50	82.50	47.50	80	91.25	81.25	98.75	100
		Z(AES)	23.75	58.75	68.75	42.50	72.50	80	62.50	95	95
		Z(AEA)	26.25	62.50	70	45	76.25	81.25	70	95	95
	NU	RL	18.75	50	72.50	90	57.50	86.25	83.75	98.75	100
		χ^2	28.75	57.50	68.75	47.50	86.25	87.50	85	97.50	98.75
		Z(AES)	2.50	3.75	6.25	0	7.50	15	2.50	12.50	23.75
		Z(AEA)	17.50	61.25	78.75	41.25	82.50	90	82.50	100	100
	M	RL	16.25	75	62.50	88.75	45	82.50	71.25	96.25	98.75
		χ^2	41.25	81.25	96.25	61.25	93.75	100	80	100	100
		Z(AES)	20	57.50	82.50	32.50	80	91.25	58.75	97.50	100
		Z(AEA)	41.25	85	100	63.75	100	100	83.75	100	100
40	U	RL	1	5	7	6	16	28	7	29	51
		χ^2	27	66	80	38	81	86	69	96	100
		Z(AES)	28	62	72	40	75	82	62	94	98
		Z(AEA)	31	69	75	42	77	83	62	94	98
	NU	RL	22	48	72	50	79	92	80	100	100
		χ^2	18	60	78	41	79	89	87	97	100
		Z(AES)	1	14	19	2	8	29	11	37	55
		Z(AEA)	18	62	85	32	82	95	82	97	100
	M	RL	16	33	47	36	79	87	75	92	99
		χ^2	42	82	95	59	97	100	86	100	100
		Z(AES)	16	47	71	29	70	86	51	83	99
		Z(AEA)	42	85	96	60	98	100	89	100	100

Tabla 3
Porcentaje de IC en la situación de no impacto (α= 0.001)

% FDI	FDI	Indice	N= 250			N= 500			N= 1.000		
			d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1
20	U	RL	0	0	2	0	10	24	0	20	34
		χ^2	16	62	72	32	76	84	78	100	100
		Z(AES)	26	50	60	30	66	68	62	94	100
		Z(AEA)	24	52	60	30	64	68	70	94	100
	NU	RL	16	42	66	56	84	84	84	100	100
		χ^2	20	34	50	58	80	92	90	100	100
		Z(AES)	0	0	0	0	0	2	0	0	0
		Z(AEA)	4	20	60	28	66	94	82	98	100
	M	RL	0	28	34	28	70	86	60	100	100
		χ^2	28	66	92	42	92	100	66	100	100
		Z(AES)	20	42	84	20	66	88	36	90	100
		Z(AEA)	14	64	96	44	92	100	74	100	100
33	U	RL	0	0	0	0	3.75	15	1.25	15	28.75
		χ^2	16.25	50	71.25	25	67.50	80	62.50	95	98.75
		Z(AES)	17.50	40	62.50	21.25	50	67.50	45	85	86.25
		Z(AEA)	18.75	40	62.50	21.25	51.25	68.75	50	85	86.25
	NU	RL	8.75	30	53.75	35	68.75	72.50	71.25	95	100
		χ^2	10	32.5	53.75	28.75	68.75	81.25	72.50	93.75	95
		Z(AES)	0	1.25	1.25	0	1.25	3.75	0	3.75	15
		Z(AEA)	3.75	21.25	51.25	12.50	57.50	77.50	61.25	95	97.50
	M	RL	2.75	11.25	26.25	20	56.25	66.25	52.50	90	97.50
		χ^2	22.50	63.75	91.25	52.50	86.25	100	75	100	100
		Z(AES)	7.50	36.25	75	20	58.75	81.25	38.75	93.75	97.50
		Z(AEA)	17.50	62.50	96.25	43.75	90	100	76.25	100	100
40	U	RL	0	0	1	1	3	13	0	5	30
		χ^2	12	54	74	24	67	80	52	92	95
		Z(AES)	12	48	68	24	62	71	40	84	87
		Z(AEA)	16	50	69	26	65	70	43	85	87
	NU	RL	4	23	49	35	61	77	58	95	99
		χ^2	6	25	56	22	58	79	71	95	98
		Z(AES)	0	3	7	0	2	13	2	17	38
		Z(AEA)	4	20	52	5	59	80	52	94	99
	M	RL	5	13	22	21	49	59	56	89	97
		χ^2	30	75	92	46	88	99	76	99	100
		Z(AES)	8	27	53	15	52	71	36	73	92
		Z(AEA)	20	65	89	38	87	98	77	99	100

Tabla 4
Porcentaje de falsos positivos. Situación de no impacto

α	% FDI	Indice	N= 250			N= 500			N= 1.000		
			d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1
.05	20%	RL	5.17	5.83	6.17	6.17	9.17	12.83	7.83	13.33	19.00
		χ^2	5.33	8.67	13.67	6.00	13.67	24.50	14.17	27.83	46.33
		Z(AES)	7.00	9.83	15.83	4.83	15.17	26.83	16.17	30.33	47.33
		Z(AEA)	11.67	15.17	20.63	11.50	21.67	34.50	23.67	37.67	54.50
	33%	RL	5.49	9.02	10.98	8.82	14.90	21.37	12.75	24.71	37.45
		χ^2	6.47	15.49	29.02	12.35	24.51	47.06	24.12	56.67	92.16
		Z(AES)	8.63	20.59	36.08	13.53	29.22	47.45	28.24	53.53	71.37
		Z(AEA)	13.92	27.84	41.18	20.59	38.04	55.29	37.05	64.51	77.03
	40%	RL	6.89	11.11	12.22	10.44	20.22	29.56	14.89	31.11	48.00
		χ^2	10.00	24.66	44.44	13.40	38.67	64.67	36.89	70.44	86.67
		Z(AES)	13.78	28.00	47.55	18.20	42.22	66.22	38.22	68.00	83.33
		Z(AEA)	19.11	36.89	55.33	24.60	52.44	72.00	44.89	76.22	85.11
0.01	20%	RL	1.17	1.33	1.83	1	2.83	4.17	1.83	3.67	6.33
		χ^2	1.33	2.50	5.33	1.17	4.50	9.67	4.17	14.50	28
		Z(AES)	1.17	3.00	7.83	1.50	5.50	12.67	4.83	17.33	29.83
		Z(AEA)	2.33	5.17	9.50	3.50	8.33	16.33	6.83	19.50	34.33
	33%	RL	1.57	3.53	1.96	1.17	4.51	5.69	3.53	9.02	16.86
		χ^2	2.16	6.09	15.49	4.31	11.57	28.43	11.76	36.08	79.02
		Z(AES)	2.17	7.45	18.23	4.71	13.33	30.59	15.10	35.29	56.86
		Z(AEA)	3.92	7.25	20.20	7.06	17.06	37.45	17.65	42.94	64.71
	40%	RL	1.78	2.89	2.89	2.67	5.56	10.44	3.78	12.89	22.89
		χ^2	1.78	9.33	27.33	5.80	20.89	46.22	21.11	54	76.89
		Z(AES)	2.67	13.55	30.89	6.40	21.78	48.22	24	53.56	72.89
		Z(AEA)	5.11	17.55	35.55	9.40	28.67	55.11	28.44	58.59	78.67
0.001	20%	RL	0	0	0.17	0	0.33	0.17	0	0.33	0.67
		χ^2	0.16	0.67	0.67	0	1.17	2.67	0.67	5.83	12.50
		Z(AES)	0.33	0.67	1.50	0.17	1.50	4.17	1.67	7.17	14.17
		Z(AEA)	0.17	0.83	1.83	0.50	2.33	5.17	2.33	7.33	16.00
	33%	RL	0.20	0.59	0	0	0.39	0.59	0.78	1.37	4.51
		χ^2	0.39	1.76	4.51	0.20	3.33	11.37	3.14	18.24	45.29
		Z(AES)	0.59	1.57	6.67	0.20	5.49	14.70	6.08	20.59	41.17
		Z(AEA)	0.78	2.16	7.25	1.18	6.86	16.86	6.86	22.16	47.27
	40%	RL	0	0.22	0.67	0.22	1.33	2.00	0.22	2.22	8.22
		χ^2	0.44	2	11.33	1.20	7.11	27.33	7.55	37.33	63.78
		Z(AES)	0.44	3.78	14.44	2.60	10.44	28.67	10.00	34.89	58.44
		Z(AEA)	0.89	4	15.56	3.00	14.00	35.11	10.89	40.67	64.22

En cuanto al tamaño muestral, se encontró que un aumento del mismo supone también un aumento tanto de la tasa de IC como de FP. El número de FP se mantuvo cerca de los niveles nominales en las condiciones menos extremas (N=250 y d_{R-F}=0.4) y fueron algo más elevados cuando los tamaños muestrales fueron altos (N=500 y N=1000), la cantidad de FDI fue mayor (0.7 y 1) y el porcentaje de items con FDI aumentó. Por el contrario, el aumento del número de items con FDI en el test no provocó una mejora en la tasa de IC, mostrándose en algunas situaciones la tendencia contraria.

En cualquier caso, tanto la tasa de IC como de FP se ven afectadas por el nivel de significación fijado. Si se considera el porcentaje de IC se observa que, en todas las condiciones de menor tamaño muestral (N=250 y N=500), sobre todo cuando la cantidad de FDI fue menor, éste disminuye conforme el nivel de significación es más restrictivo. Sin embargo, en las condiciones de tamaño muestral mayor ésta tendencia no se presenta tan marcada, de tal modo que los items correctamente identificados al 5% también lo son a niveles de significación más bajos (1% y 0.1%), excepto en el caso de FDI uniforme. El porcentaje de FP disminuye con el incremento del nivel de confianza, es decir, cuando los niveles de significación considerados fueron los más extremos, la tasa de FP fue también muy baja.

Medidas de Raju y de Lord. Los resultados encontrados, cuando se aplicaron las medidas de área de Raju y el estadístico de Lord, muestran que tanto el estadístico de Lord como Z(AEA) alcanzaron porcentajes de IC altos y similares. La medida de exacta de área con signo (Z(AES)) fue incapaz de detectar FDI no uniforme. Sin embargo, en cuanto a la identificación del FDI mixto y uniforme las tasas de IC para dicha medida con signo, concuerdan con las encontradas en los otros dos estadísticos utilizados (χ^2 y Z(AEA)) (cf. tablas 1, 2 y 3). Al aumentar el número de items con FDI en el test mejoró la capacidad de Z(AES) para detectar correctamente items con FDI uniforme. Tal y como era de esperar a mayor cantidad de FDI generado mayor porcentaje de IC, tendencia que se mantiene a través de los diferentes tamaños muestrales y en las distintas pruebas estadísticas comparadas.

En cuanto al tamaño muestral, conforme éste aumenta se incrementa tanto la tasa de IC como de FP en los tres índices bajo estudio. El número de FP (cf. tabla 4) incrementó con el aumento del tamaño muestral, cantidad de FDI y porcentaje de items con FDI en el test. Solamente se mantienen cerca de los niveles nominales cuando la cantidad de FDI fue menor, el tamaño muestral fue de 250 ó de 500 sujetos y el porcentaje de items con FDI fue del 20%.

Cuando N=1000 y d_{R-F}=1, el porcentaje de FP fue muy alto, más de la mitad de los items incorrectamente identificados, porcentaje que fue en aumento al incrementarse el número de items con FDI en el test. El estadístico de Lord mostró las tasas de FP más bajas, seguido de Z(AES) y de Z(AEA), sin embargo, con el incremento del porcentaje de items con FDI y en las condiciones más extremas de tamaño muestral y cantidad de FDI, el estadístico de Lord presentó tasas de FP ligeramente superiores a las obtenidas con los estadísticos de Raju.

La tasa de IC como de FP también estuvieron afectadas por el nivel de significación fijado. Si se considera el porcentaje de IC, se observa que en todas las condiciones de tamaño muestral pequeño y sobre todo cuando la cantidad de FDI fue menor, éste disminuye conforme el nivel de significación es más restrictivo. Sin embargo, en las condiciones de tamaño muestral mayor, ésta tendencia no se presenta tan marcada, de tal modo que los items correctamente identificados al 5%, también lo son a niveles de significación más bajos (1% y 0.1%). Por

otro lado, con relación a la tasa de FP, ésta disminuye en todos los tamaños muestrales y condiciones con el incremento del nivel de confianza, es decir, cuando los niveles de significación considerados fueron los más bajos, la tasa de FP fue también muy baja. Aún así, en tamaños muestrales elevados y cuando la cantidad de FDI fue mayor, el número de FP fue considerable, incluso al nivel de significación más bajo.

Situación de Impacto

Regresión Logística. En las tablas 5 a la 8 aparecen los resultados encontrados en la situación de impacto cuando se aplicó análisis de RL. Se observa que, independientemente del tipo de FDI y del tamaño muestral de los grupos focal y de referencia, conforme aumenta la cantidad de FDI generado también aumenta el número de IC. Sin embargo, este incremento fue mayor de la condición 1 (d_{R-F}=0.4) a la condición 2 (d_{R-F}=0.7), que de ésta última a la condición 3 (d_{R-F}=1).

Si se considera el tipo de FDI, tamaño muestral y nivel de significación se observa que el comportamiento del procedimiento de RL fue idéntico al obtenido en la situación de no impacto tanto con respecto al porcentaje de IC como al de FP.

Medidas de Raju y de Lord. Los resultados encontrados cuando se utilizaron las medidas de área y el estadístico de Lord (cf. tablas 5, 6 y 7) muestran que conforme aumenta la cantidad de FDI generado también aumenta el número de IC, tendencia que se mantiene en los tamaños muestrales de 250 y 500 sujetos. Cuando N=1000, la tasa de IC fue del 100% en χ^2 y Z(AEA) independientemente de la cantidad de FDI generado y del porcentaje de items con FDI. Por otro lado, Z(AES) alcanzó menor porcentaje de IC en relación a los otros dos estadísticos, y detectó peor el FDI no uniforme frente al FDI uniforme y mixto. Los estadísticos χ^2 y Z(AEA), siguiendo la pauta encontrada en la situación de no impacto, presentan porcentajes de IC similares en todas los tipos de FDI estudiados. No obstante, estos estadísticos identificaron menor porcentaje de items con FDI uniforme que con FDI no uniforme o mixto. Estos resultados fueron mejores con el aumento del tamaño muestral.

El número de FP se incrementó con el aumento en tamaño muestral y cantidad de FDI. La tasa de FP también estuvo afectada por el porcentaje de items con FDI, cuando éste aumentó también aumentaron el número de FP. Esto se produjo principalmente en las situaciones de menor tamaño muestral. En contraposición, el número de IC no varió con dicho aumento, excepto en la medida Z(AES) que mejoró la identificación de FDI no uniforme.

Las tasas de IC y FP estuvieron afectadas por el nivel de significación. En N=250 (independientemente de la cantidad de FDI generado) y N=500 (condición de d_{R-F}=0.4) se observa que el porcentaje de IC disminuye conforme el nivel de significación se hace más restrictivo. En N=500 (condiciones de d_{R-F}=0.7 y d_{R-F}=1) y N=1000 (en todas las condiciones de cantidad de FDI) todos los items son identificados correctamente al 5% y también al 0.01%. En cuanto a la tasa de FP, ésta disminuye al considerar niveles de significación más restrictivos solamente en las condiciones de menor tamaño muestral.

Discusión

El porcentaje de identificaciones correctas, tal y como era de esperar, se dejo afectar por el tamaño muestral y la cantidad de FDI generado. Así, independientemente del estadístico de evaluación del FDI considerado, en las condiciones de mayor tamaño muestral y cantidad de FDI se produjeron mayor número de identificaciones correctas. Por el contrario, la presencia de diferencias entre grupos (impacto) no pareció afectar la precisión en la correcta identificación de items

Tabla 5
Porcentaje de IC en la situación de impacto (α= 0.05)

% FDI	FDI	Indice	N= 250			N= 500			N= 1.000		
			d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1
20	U	RL	14	20	38	24	44	44	28	50	60
		χ^2	42	66	62	58	66	66	100	100	98
		Z(AES)	44	66	60	54	60	60	84	98	98
		Z(AEA)	46	68	62	58	60	60	100	100	100
	NU	RL	66	90	92	90	100	100	96	100	100
		χ^2	68	86	88	84	94	100	100	100	100
		Z(AES)	0	8	12	8	6	16	74	76	78
		Z(AEA)	60	98	96	84	100	98	100	100	100
	M	RL	52	64	62	76	94	84	100	100	92
		χ^2	50	94	100	78	100	100	100	100	98
		Z(AES)	18	70	78	46	82	82	78	92	90
		Z(AEA)	52	98	98	94	100	100	100	100	100
33	U	RL	10.00	21.25	31.25	22.50	32.50	55.00	32.50	47.50	57.50
		χ^2	32.50	62.50	62.50	47.50	65.00	65.00	100	100	98.75
		Z(AES)	32.50	60.00	62.50	50.00	62.50	62.50	88.75	98.75	98.75
		Z(AEA)	36.25	62.50	62.50	52.50	62.50	62.50	100	100	100
	NU	RL	52.50	76.25	86.25	83.75	93.75	96.25	91.25	97.50	100
		χ^2	52.50	73.75	76.25	71.25	90.00	85.00	100	100	100
		Z(AES)	7.50	11.25	17.50	12.50	12.50	20.00	78.75	77.50	76.25
		Z(AEA)	42.50	85.00	86.25	72.50	92.50	91.25	100	100	100
	M	RL	43.75	55.00	56.25	63.75	67.50	66.25	93.75	97.50	92.50
		χ^2	56.25	95.00	95.00	73.75	100	92.50	100	100	98.75
		Z(AES)	23.75	57.50	73.75	46.25	73.75	78.75	77.50	91.25	97.50
		Z(AEA)	63.75	86.25	91.25	86.25	93.75	95.00	100	100	100
40	U	RL	10	20	29	21	33	50	33	54	68
		χ^2	34	68	71	50	69	72	100	99	99
		Z(AES)	37	70	70	49	68	70	93	98	100
		Z(AEA)	40	71	70	56	69	70	100	99	100
	NU	RL	48	77	89	78	91	93	91	99	100
		χ^2	48	78	87	64	88	91	100	100	100
		Z(AES)	14	22	29	10	32	38	82	81	87
		Z(AEA)	55	81	93	70	92	94	100	100	100
	M	RL	39	45	43	65	71	61	78	86	72
		χ^2	65	90	88	82	97	97	100	100	96
		Z(AES)	18	45	58	27	59	59	86	95	93
		Z(AEA)	61	83	83	78	87	85	100	100	100

Tabla 6
Porcentaje de IC en la situación de impacto (α= 0.01)

% FDI	FDI	Indice	N= 250			N= 500			N= 1.000		
			d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1
20	U	RL	8	12	22	14	25	36	18	38	50
		χ^2	22	64	60	48	60	60	100	100	98
		Z(AES)	20	56	60	48	60	60	80	88	98
		Z(AEA)	24	58	60	54	60	60	100	100	98
	NU	RL	32	72	78	72	94	98	90	100	100
		χ^2	42	76	72	72	92	98	100	100	100
		Z(AES)	0	6	2	2	6	8	62	62	64
		Z(AEA)	36	78	86	54	94	96	100	100	100
	M	RL	12	40	44	52	78	64	92	100	88
		χ^2	36	84	94	60	100	100	100	100	96
		Z(AES)	4	56	78	30	74	80	70	82	82
		Z(AEA)	36	90	96	54	98	100	100	100	100
33	U	RL	1.25	8.75	18.75	7.50	15	27.50	23.75	36.25	50
		χ^2	21.25	58.75	61.25	35	61.25	62.50	100	100	98.75
		Z(AES)	20	52.50	61.25	23.75	62.50	62.50	87.50	93.75	97.50
		Z(AEA)	23.75	55	61.25	27.50	61.25	62.50	100	100	98.75
	NU	RL	36.25	65	73.75	68.75	88.75	87.50	87.50	90	96.25
		χ^2	26.25	56.25	65	53.75	80	80	100	100	100
		Z(AES)	3.75	1.25	3.75	1.25	5	8.75	70	67.50	68.75
		Z(AEA)	18.75	65	77.50	43.75	88.75	85	100	100	100
	M	RL	10	28.75	32.50	43.75	55	48.75	71.25	91.25	80
		χ^2	41.25	78.75	90	60	97.50	91.25	100	100	96.25
		Z(AES)	12.50	38.75	70	18.75	66.25	71.25	70	82.50	90
		Z(AEA)	35	73.75	88.75	55	90	88.75	100	100	100
40	U	RL	4	10	15	9	18	35	21	37	53
		χ^2	20	56	70	31	68	71	100	99	99
		Z(AES)	22	58	69	33	68	70	91	93	99
		Z(AEA)	23	59	69	36	68	70	100	99	100
	NU	RL	34	66	73	69	91	89	87	94	100
		χ^2	28	59	69	45	76	85	100	100	100
		Z(AES)	4	11	13	3	18	26	73	69	80
		Z(AEA)	21	66	80	45	80	91	100	100	100
	M	RL	13	29	28	42	50	43	68	78	66
		χ^2	43	77	81	61	95	92	100	100	95
		Z(AES)	7	32	50	12	48	56	77	87	87
		Z(AEA)	28	69	78	54	83	83	100	100	100

Tabla 7
Porcentaje de IC en la situación de impacto (α= 0.001)

% FDI	FDI	Indice	N= 250			N= 500			N= 1.000		
			d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1
20	U	RL	0	0	12	6	16	22	8	22	38
		χ^2	10	46	58	34	60	60	100	100	96
		Z(AES)	6	44	60	36	56	60	80	84	84
		Z(AEA)	6	46	60	36	56	60	100	100	96
	NU	RL	26	42	50	46	82	92	76	98	98
		χ^2	18	36	38	46	62	76	100	100	100
		Z(AES)	0	0	0	0	0	0	60	60	60
		Z(AEA)	10	36	44	22	64	90	100	100	100
	M	RL	4	18	22	30	56	56	74	94	84
		χ^2	24	66	94	44	94	98	100	100	90
		Z(AES)	0	28	68	6	60	74	68	76	70
		Z(AEA)	10	54	94	30	94	98	100	100	100
33	U	RL	0	2.50	6.25	3.75	6.25	13.75	13.75	18.75	42.50
		χ^2	1.25	38.75	53.75	12.50	52.50	62.50	100	100	98.75
		Z(AES)	7.50	33.75	55	15	52.50	57.50	87.50	90	92.50
		Z(AEA)	7.50	35	55	16.25	52.50	58.75	100	100	100
	NU	RL	17.50	42.50	55	43.75	72.50	83.75	72.50	87.50	93.75
		χ^2	8.75	31.25	45	30	67.50	57.50	100	100	100
		Z(AES)	1.25	0	2.50	0	2.50	2.50	62.50	62.50	62.50
		Z(AEA)	6.25	28.75	53.75	18.75	56.25	73.75	100	100	100
	M	RL	2.50	7.50	13.75	20	38.75	42.50	55	70	63.75
		χ^2	22.50	56.25	83.75	45	93.75	85	100	100	93.75
		Z(AES)	3.75	16.25	52.50	3.75	51.25	66.25	70	72.50	78.75
		Z(AEA)	13.75	43.75	78.75	26.25	86.25	86.25	100	100	100
40	U	RL	1	3	4	5	11	20	10	22	34
		χ^2	4	38	59	9	62	69	100	99	99
		Z(AES)	7	37	57	12	61	68	90	89	97
		Z(AEA)	6	37	59	12	61	68	100	99	99
	NU	RL	18	37	55	44	76	87	75	90	97
		χ^2	9	33	35	30	58	65	100	100	99
		Z(AES)	2	6	7	1	8	10	68	68	74
		Z(AEA)	7	31	53	19	58	84	100	100	100
	M	RL	1	13	16	23	34	35	42	56	49
		χ^2	14	52	72	38	88	83	100	100	88
		Z(AES)	0	9	31	5	28	49	76	78	71
		Z(AEA)	5	37	63	26	73	78	100	100	100

Tabla 8
Porcentaje de falsos positivos. Situación de impacto

α	% FDI	Indice	N= 250			N= 500			N= 1.000		
			d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1	d= 0.4	d= 0.7	d= 1
0.05	20%	RL	4.83	6.33	5.83	7.67	6.33	7.50	9.00	8.67	10.00
		χ^2	6.17	10.33	12.00	7.67	14.17	24.33	100	100	100
		Z(AES)	6.83	10.67	12.33	8.17	16.33	24.17	96.50	97.67	96.17
		Z(AEA)	11.17	17.33	19.83	11.83	22.50	31.50	100	100	100
	33%	RL	4.17	7.06	7.06	5.29	7.25	11.37	7.25	12.55	17.45
		χ^2	8.63	17.25	31.18	11.96	30.98	44.20	100	100	100
		Z(AES)	10.00	18.23	34.31	12.75	33.53	42.60	98.24	98.63	99.80
		Z(AEA)	15.88	26.47	41.76	18.82	40.98	51.80	100	100	100
	40%	RL	5.11	7.11	9.11	5.56	10.44	14.00	9.56	16.22	22.00
		χ^2	11.33	27.11	41.78	20.22	45.11	62.44	100	100	100
		Z(AES)	12.00	27.33	44.44	20.22	40.67	58.00	100	99.78	100
		Z(AEA)	18.67	38.44	51.33	29.56	51.56	64.22	100	100	100
0.01	20%	RL	1.33	1.33	0.67	1.17	1.00	1.83	2.33	2	2.83
		χ^2	2.33	2	4.33	2.17	5.00	9.50	100	100	100
		Z(AES)	3.17	3.17	5.50	2.50	5.33	11.17	95.67	96.33	95.50
		Z(AEA)	4.17	4.33	7.67	4.33	8.67	14.00	100	100	100
	33%	RL	1.37	1.76	1.76	0.98	1.76	2.35	1.57	3.92	6.27
		χ^2	2.55	5.10	15.10	3.92	14.31	27.80	100	100	100
		Z(AES)	3.14	8.04	19.21	4.12	17.65	29.40	96.86	97.25	98.62
		Z(AEA)	4.12	11.18	25.29	6.47	21.57	35.40	100	100	100
	40%	RL	1.33	2.22	1.33	1.11	2.89	5.78	2.89	6	7.56
		χ^2	5.78	12.89	22.67	7.33	24.00	44.00	100	100	100
		Z(AES)	6.00	14.00	25.78	8.67	23.57	42.22	98.89	98.67	99.55
		Z(AEA)	8.00	19.79	31.33	12.67	29.79	49.11	100	100	100
0.001	20%	RL	0	0	0	0.17	0.33	0.17	0.33	0.50	0.33
		χ^2	0.50	0.33	1.17	0.50	1.00	4.33	100.00	100.00	100.00
		Z(AES)	0.50	0.33	1.50	0.83	1.50	4.50	94.00	94.50	94.17
		Z(AEA)	0.67	0.67	2.00	1.00	2.33	5.00	100.00	100.00	100.00
	33%	RL	0.20	0	0.20	0.20	0.39	0.00	0.20	0.78	0.59
		χ^2	0.20	1.57	3.73	0.78	4.71	12.80	100.00	100.00	100.00
		Z(AES)	0.78	2.75	6.86	1.37	5.88	15.20	95.88	96.08	96.67
		Z(AEA)	0.59	2.75	9.22	1.76	6.47	17.20	100.00	100.00	100.00
	40%	RL	0	0	0.22	0	0.44	0.67	0.44	1.33	0.89
		χ^2	0.44	5.11	9.55	2.67	10.44	24.44	100.00	100.00	100.00
		Z(AES)	1.56	5.33	13.56	3.33	12.00	24.89	97.11	96.44	97.55
		Z(AEA)	2.44	6.67	14.89	3.78	14.00	28.22	100.00	100.00	100.00

con FDI, algo a esperar dado que tanto RL como los estadísticos de la TRI son métodos que evalúan la presencia de FDI en función del nivel de habilidad. Por otro lado, la cantidad de items con FDI en el test tampoco resultó relevante en la detección correcta de items que funcionan diferencialmente.

Los estadísticos de la TRI frente al análisis de RL identificaron mejor el funcionamiento diferencial uniforme. Por contra, sólo en las condiciones menos extremas (menor porcentaje de items con FDI en el test, menor cantidad de FDI, menor tamaño muestral) los procedimientos derivados de la TRI obtienen porcentajes de IC más altos que RL en cuanto al FDI mixto. Por último, el FDI no uniforme es detectado igualmente bien por RL, Z(AEA) y χ^2 de Lord, siendo Z(AES) la que obtuvo los porcentajes de IC más bajos. El estadístico de Lord fue más efectivo en la identificación de FDI que Z(AES) y Z(AEA), resultados que concuerdan con los encontrados por Cohen y Kim (1993). Estos resultados llevan, en principio, a preferir el estadístico de Lord sobre el resto de procedimientos utilizados. Sin embargo, tanto χ^2 de Lord como Z(AEA) y Z(AES) controlan peor el porcentaje de FP, de tal modo que los valores encontrados fueron muy elevados y superiores a los encontrados cuando se utilizó RL. Esta circunstancia se dio principalmente en las condiciones de mayor cantidad de FDI, mayor tamaño muestral, mayor porcentaje de items con FDI y presencia de impacto entre grupos. En el caso de los estadísticos de la TRI, la presencia en el test de un alto porcentaje de items con FDI, las diferencias entre grupos y entre parámetros, pueden estar afectando seriamente al cálculo de las constantes de igualación. En este sentido, el cálculo de las mismas puede ser incorrecto y enmascarar la identificación correcta de FDI al mismo tiempo que identificaría un gran número de falsos positivos. Bajo estas situaciones es mejor utilizar un procedimiento iterativo de igualación (Candell y Drasgow, 1988; Kim y Cohen, 1992a; Lautenschlager, Flaherty y Park, 1994; Lautenschlager y Park, 1988; Miller y Oshima, 1992; Park y Lautenschlager, 1990) dado que proporciona resultados más fiables y decrementa el número de FP (Kim y Cohen, 1992a). En este punto, resultaría interesante comparar los efectos que se producirían al utilizar un procedimiento iterativo de purificación de la habilidad tanto para RL como los estadísticos de Lord y de Raju.

De los resultados obtenidos en este estudio se deduce también que, cuando se empleen las medidas de área de Raju y el estadístico de Lord, se debe trabajar con niveles de significación del 1% ó del 0.1%, dado que el porcentaje de FP se reduce sin disminución del porcentaje de IC.

En resumen, resulta más aconsejable, a la vista de los resultados aportados en este trabajo, el estadístico de Lord sobre el resto de procedimientos estudiados. Sin embargo, debido al escaso control que éste ejerce sobre la tasa de FP se recomienda emplear este estadístico junto a otra/s medidas de evaluación del FDI. Cohen y Kim (1993) sugieren que conjuntamente al estadístico de Lord se calculen las medidas exactas de área de Raju, la utilización de ambos procedimientos proporcionaría información complementaria en la evaluación del FDI. Los resultados de éste estudio apuntan a que el análisis de RL también puede utilizarse junto al estadístico de Lord. El análisis de RL fue precisamente el procedimiento que en las condiciones más extremas mostró el menor porcentaje de FP, si bien presentó el peor porcentaje de IC en lo que se refirió a la identificación de FDI uniforme.

Agradecimientos

Los autores agradecen al editor y a un revisor anónimo sus valiosos comentarios sobre la primera versión de este manuscrito.

Referencias

Baker, F.B. (1993). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. [Computer program] Madison WI: University of Wisconsin. Laboratory of Experimental Design.

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

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

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.

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. (1996). Funcionamiento diferencial de los items. Procedimiento Mantel-Haenszel y modelos loglineales. Tesis doctoral no publicada. Universidad de Oviedo.

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. (1995). Evaluación del funcionamiento diferencial del ítem en items dicotómicos y politómicos: un estudio comparativo. Tesis doctoral no publicada. Murcia, Universidad de Murcia.

Hidalgo, M.D. y López Pina, 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 Ciencias del Comportamiento, La Manga, Murcia.

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, N.J.: Erlbaum.

Hosmer, D.W. y Lemeshow, S. (1989). Applied Logistic Regression. New York, NY: Wiley.

Kim, S.H. y Cohen, A.S. (1991). A comparison of two area measures for detecting differential item functioning. Applied Psychological Measurement, 15, 269-278.

Kim, S.H. y Cohen, A.S. (1992a). Effects of linking methods on detection of DIF. Journal of Educational Measurement, 29, 51-66.

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

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, N.J.: Erlbaum.

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

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

Millsap, R.E. y Everson, H.T. (1993). Methodology Review: Statistical Approaches for Assesing Measurement Bias. Applied Psychological Measurement, 17, 297-334.

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

Navas, M.J. y Gómez, J. (1994). Comparison of several bias detection techniques. Paper presented at the 23rd. International Congress of Applied Psychology, Madrid.

Park, D.G. 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.

Rogers, H.J. y Swaminathan, H. (1993). A comparison of Logistic Regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, 17, 105-116.

Scheuneman, J. (1979). A new method for assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Steinberg, D. y Phillips, C. (1991). LOGIT: A supplementary module for SYSTAT. Evanston, IL: SYSTAT, Inc.

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

Swaminathan, H. y Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.

Thissen, D., Steinberg, L. y Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. En H. Wainer y H.I. Braun (Eds.) Test Validity. Hillsdale, N.J.: Erlbaum.

Thissen, D., Steinberg, L. y Wainer, H. (1993). Detection of Differential Item Functioning Using the Parameters of Item Response Models. En P.W. Holland y H. Wainer (Eds.) Differential Item Functioning (pp. 67-113). Hillsdale, NJ: LEA.

Aceptado el 8 de octubre de 1996