Psicothema
2000. Vol. 12, Supl. nº 2, pp. 399-402

ISSN 0214 - 9915 CODEN PSOTEG

Comparación de dos procedimientos de purificación del test para la evaluación del FDI con el estadístico de Lord y con las medidas de área de Raju

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

Diversos estudios han comprobado la efectividad de un proceso de purificación iterativo del test para mejorar la identificación del Funcionamiento Diferencial del Ítem (FDI) frente a un proceso no iterativo. Sin embargo, los procedimientos iterativos son muy costosos, ya que implican una elevada cantidad de cálculos. En este estudio se comparan los procedimientos de purificación bietápicos de Miller y Oshima (1992) e Hidalgo y López (1997), tomando como índices las tasas de identificaciones correctas y las tasas de falsos positivos, usando como medidas de evaluación del FDI el estadístico de Lord (1980) y las medidas de área exacta de Raju (1990).

Comparison of two test purification procedures to assess DIF with Lord's statistic and Raju's area measures. Several studies have tested the effectiviteness of iterative test purification procedure in order to improve the identification of Differential Item Functioning (DIF) againts no-iterative test purification procedures. However, iterative test purification procedures are very cost because they have many calculations. In this study the bietapic purification prodedures of Miller & Oshima (1992) and bietapic purification procedure of Hidalgo & López (1997) are compared, by taking into account correct identifications and false positives rates, and carrying out Lord's statistic (1980) and Raju's exact area measures (1990) like DIF assessing measures.

El *Funcionamiento Diferencial del Ítem (FDI)* define al ítem de un test que presenta diferencias estadísticas cuando es contestado en distintos grupos, los cuales han sido igualados previamente en un nivel de habilidad. Si un ítem funciona diferencialmente se está violando uno de los supuestos fundamentales de la TRI, el supuesto de invarianza de los parámetros del ítem.

Se han propuesto diversos estadísticos y procedimientos para detectar ítems con FDI basados en los supuestos de la TRI (Gómez e Hidalgo, 1997; Millsap y Everson, 1993). Antes de usar alguno de estos procedimientos, el primer paso es estimar los parámetros de los ítems y de la habilidad en cada uno de los grupos. A continuación, para poder compararlos, hay que transformar los parámetros a una métrica común mediante un método de enlace y, una vez igualados los grupos, se pueden calcular los índices de evaluación del FDI. Sin embargo, la presencia de ítems con FDI en el test afecta a los procesos de estimación de los parámetros y al cálculo de las constantes de igualación, lo cual falseará los resultados de las identificaciones de ítems con FDI en el test. Para superar este problema se han propuesto diversos *procedimientos de purificación del test* (Candell y Drasgow, 1988; Hidalgo y López, 1997; Lautenschlager, Flaherty y Park, 1994; Lord, 1980; Miller y Oshima, 1992; Park y Lautenschlager, 1990). Todos ellos son procedimientos iterativos basados en la reestimación y reigualación de los parámetros. Los procedimientos de Candell y Drasgow (1988), Lautenschlager et al. (1994), Lord (1980) y Park y Lautenschlager (1990) siguen un proceso iterativo que conlleva un elevado costo computacional y temporal, por lo que han sido simplificados en el número de iteraciones por los procedimientos bietápicos de Hidalgo y López (1997) y de Miller y Oshima (1992).

El procedimiento de Miller y Oshima es una modificación de los procedimientos iterativos de Lautenschlager et al. (1994) y Park y Lautenschlager (1990). Se basa en la reestimación de los parámetros de los ítems (PBE). El procedimiento bietápico de Hi-

Tabla 1
Procedimientos de purificación bietápicos

Miller y Oshima (1992) Procedimiento bietápico de estimación (PBE)	**Hidalgo y López (1997)** Procedimiento bietápico de igualación (PBI)
1ª ETAPA	**1ª ETAPA**
1. Estimación de los parámetros de los ítems	1. Estimación de los parámetros de los ítems
2. Igualación de las métricas	2. Igualación de las métricas
3. Cálculo de los índices del FDI	3. Cálculo de los índices del FDI
4. Eliminar del test los ítems con FDI	4. Eliminar del test los ítems con FDI
2ª ETAPA	**2ª ETAPA**
1. Reestimar los parámetros de los ítems sin FDI	1. Reigualar las métricas
2. Igualar las métricas	2. Cálculo de los índices del FDI
3. Cálculo de los índices del FDI	

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

dalgo y López (1997) simplifica el procedimiento de purificación iterativo de Candell y Drasgow (1988), centrado en la reigualación de los grupos (PBI).

La diferencia entre estos dos procedimientos (tabla 1) está en la segunda etapa, mientras que en el procedimiento PBE se reestiman los parámetros de los ítems, en el procedimiento PBI, tras eliminar del test los ítems identificados con FDI, se reigualan las métricas de los parámetros.

Los propósitos de este estudio son:

• Estudiar el efecto de la purificación del test con los procedimientos de purificación bietápicos PBE y PBI en la detección del FDI, tomando como índices las tasas de identificaciones correctas (IC) y las tasas de falsos positivos (FP) que arrojan el estadístico χ^2 de Lord y las medidas de área con signo (Z(SA)) y sin signo (Z(H)) de Raju.

• Comparar los dos procedimientos de purificación bietápicos, teniendo como índices comparativos las tasas de identificaciones correctas (IC) y de falsos positivos (FP).

El estadístico χ^2 de Lord

Este procedimiento se basa en la comparación de los parámetros del ítem. La hipótesis nula establece la igualdad de los parámetros en dos grupos en el mismo nivel de habilidad $H_0 : a_{jF} = a_{jR}$, $b_{jF} = b_{jR}$, $c_{jR} = c_{jF}$, donde j designa al ítem, R al grupo de referencia y F al grupo focal. Un ítem no presenta FDI cuando se acepta la hipótesis nula de igualdad de los parámetros.

El estadístico χ^2 de Lord es el resultado de una operación vectorial:

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

donde,

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

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

y $\sum_j$ es la matriz de dispersión 2x2, tal que $\sum_j = \sum_{jR} + \sum_{jF}$, donde $\sum_{jR}$ y $\sum_{jF}$ 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) desarrolló las medidas de área exactas con signo (SA) y sin signo (UA) como índices de detección y evaluación FDI, para los modelos logísticos 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 tiene una distribución muestral asintótica (Raju, 1990), lo cual permite estudiar si las diferencias en las CCIs de los grupos se deben a errores aleatorios de muestreo o a diferencias estadísticamente significativas, utilizando una prueba Z. Para el modelo de 2-p, estas medidas son:

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

y

$$UA = \left|\hat{b}_{jF} - \hat{b}_{jR}\right| \; si \; \hat{a}_{jR} = \hat{a}_{jF}, \; o \; UA = \left|H_j\right| \; en \; cualquier \; otro \; caso.$$

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, $Z_j(SA)$, para la medida de área con signo se define como:

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

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 se utilice la variable H_j para probar la significación de UA. Entonces, la prueba estadística se define como:

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

Método

Se ha utilizado un test de 40 ítems de respuesta dicotómica ajustados a un modelo logístico de 2-p, cuyos parámetros de dificultad y discriminación fueron tomados del estudio de Candell y Drasgow (1988). La distribución de habilidad fue normal de media 0 y desviación típica 1, para los grupos de referencia y focal, generadas con el programa SYSTAT (Wilkinson, 1990), y un tamaño muestral de 500 sujetos para ambos grupos.

Antes de generar las matrices de respuesta, se manipularon dos cantidades de FDI, 0'4 y 0'6, sobre los parámetros iniciales, y dos porcentajes de ítems con FDI, 10% y 30% de los ítems, sobre los tres tipos de FDI (uniforme, no uniforme y mixto). Así, se obtienen cuatro condiciones experimentales. Las matrices de respuesta del grupo de referencia y los cuatro grupos focales fueron simuladas con el programa SIMULA v. 2 (Hidalgo y López, 1995), y 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.

Con el programa BILOG v. 3.04 (Mislevy y Bock, 1990) se estimaron los parámetros de los ítems de cada una de las matrices de respuesta simuladas por máxima verosimilitud marginal con estimación bayesiana. Para la igualación de métricas se recurrió al método de las curvas características de Stocking y Lord (1983) con el programa EQUATE v. 2.0 (Baker, 1993).

Para el estudio de FDI se ha recurrido al estadístico de Lord (1980) y a las medidas de área de Raju (1990) ejecutados con el programa IRTDIF (Kim y Cohen, 1992).

Resultados y conclusiones

Los porcentajes de identificaciones correctas (IC) y de falsos positivos (FP) obtenidos en un nivel de significación del 5%, aparecen en las tablas 2 y 3, respectivamente.

Tanto el procedimiento PBE como el PBI no mejoran, en líneas generales, las tasas de identificaciones correctas del estadístico de Lord y de las medidas de área exacta de Raju, aunque los porcentajes de estas identificaciones con PBE son más altos que los de PBI. El aumento del número de identificaciones correctas tras la purificación aparece en casos aislados. Por ejemplo, con PBE hay un aumento de las tasas de IC cuando el test tiene un 10% de los ítems con FDI uniforme y se emplea la medida de área sin signo de Raju (62% y 93%), y cuando el test tiene un 10% de ítems con FDI no uniforme y se emplea el estadístico de Lord (58% y 86%).

Las tasas de falsos positivos se reducen sustancialmente en ambos procedimientos en todas las condiciones experimentales. El estadístico de Lord y la medida de área con signo de Raju mantienen las tasas de falsos positivos próximas a los niveles nominales, siendo éstas más elevadas cuando el test tiene un 30% de ítems con FDI. La medida de área sin signo comete más falsos positivos, algunos de ellos algo alejados del nivel nominal, por ejemplo, cuando el test tiene un 30% de ítems con FDI en una magnitud de 0'6, con PBI hay un 11% de falsos positivos, y con PBE un 10'29% de falsos positivos.

Si comparamos ambos procesos de purificación bietápicos, los resultados sugieren que reestimar los parámetros es una medida conveniente, ya que en general, las tasas de identificaciones correctas al término de PBE superiores a las obtenidas con PBI, aunque ninguno de los dos produzca mejorías frente a las tasas de IC de la primera etapa, como se apuntaba más arriba. La reestimación de parámetros también consigue reducir las tasas de falsos positivos en el estadístico de Lord y en la medida de área con signo, salvo alguna excepción en esta medida (sí el test tiene un 10% de ítems con FDI con una magnitud de 0'4). Las tasas de falsos positivos en la medida de área sin signo son más elevadas con PBE que con PBI.

En definitiva:

• Ninguno de los dos procedimientos de purificación produce mejorías en las tasas de identificaciones correctas.

• Ambos procedimientos reducen sustancialmente las tasas de falsos positivos.

• La reducción de falsos positivos es ligeramente mayor cuando se utilizó el procedimiento de Miller y Oshima en las condiciones más adversas de alto porcentaje de ítems con FDI (30%) en mayor cantidad (0'6).

• En general, las diferencias entre ambos procedimientos no son acusadas aunque parece aconsejable emplear el procedimiento de purificación de Miller y Oshima (1992), sin embargo, el procedimiento de purificación de Hidalgo y López (1997) supone un ahorro computacional significante.

Tabla 2
Tasas de identificaciones correctas

			χ^2			Z(SA)			Z(H)		
			1ª	PBE	PBI	1ª	PBE	PBI	1ª	PBE	PBI
Uniforme	**0'4**	**10%**	52	49	48	58	53	50	59	**62**	52
		30%	76	59	55	75'5	66'5	53'5	81	65'5	60'5
	0'6	**10%**	93	90	88	85	84	82	91	**93**	90
		30%	98'5	88'5	87'5	98	87'5	82	99	89'5	88'5
No	**0'4**	**10%**	56	**58**	**56**	14	12	10	64	44	62
Uniforme		**30%**	68	57'5	55	6	3'5	3'5	65	**72**	48'5
	0'6	**10%**	82	**86**	80	4	—	2	90	76	82
		30%	92'5	**93**	92	11	3'5	5	92	91	89
Mixto	**0'4**	**10%**	62	52	54	60	**60**	58	76	74	66
		30%	95	85	80'5	95'5	85'5	82'5	97	91'5	87'5
	0'6	**10%**	98	96	94	96	90	88	100	**100**	98
		30%	100	99'5	98'5	100	**100**	**100**	100	99'5	99'5

Tabla 3
Tasas de falsos positivos

		χ^2			Z(SA)			Z(H)		
		1ª	PBE	PBI	1ª	PBE	PBI	1ª	PBE	PBI
0'4	**10%**	5'61	4'22	4'22	6'56	4'67	5'06	8'72	8'5	7'11
	30%	8'5	5	6'07	8'5	5'64	5'36	12'5	9'71	9'64
0'6	**10%**	6'17	5'06	5'72	7'89	5'11	5'72	10'78	9	8'33
	30%	13'07	5'07	6'64	15'57	6'71	8	18	10'29	11

Referencias

Baker, F.B. (1993). EQUATE v. 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.

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

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. (1997). *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.

Kim, S.H. y Cohen, A.S. (1992). 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.

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

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 mesurement bias. *Applied Psychological Measurement, 16*, 297-334.

Mislevy, R.J. y Bock, R.D. (1990). *PC-BILOG 3.04: 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.

Wilkinson, L. (1990). *SYSTAT: The system for Statistics (versión 5).* Evanston, IL: Systat Corporation.