Psicothema 2005. Vol. 17, nº 2, pp. 350-355
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Detección errónea del funcionamiento diferencial del ítem. Una comparación de métodos

María Ester Aguerri, María Silvia Galibert, Marta L. Zanelli* y Horacio F. Attorresi
Universidad de Buenos Aires e * Instituto Nacional de Tecnología Agropecuaria

Se comparan distintos métodos para analizar el Funcionamiento Diferencial del Ítem (DIF), en términos de sus proporciones de falsos positivos. Éstos son: la prueba normal para la diferencia de los parámetros de dificultad (PN), la prueba χ^2 de Mantel-Haenszel y la clasificación en 'tipo C' o no según la sola magnitud del estadístico MH D-DIF (clasificación empírica) —estos últimos de manera estándar y bietápica— y según el criterio del ETS. Se aplicaron a datos simulados sin DIF bajo distintas condiciones de diseño según tamaño de muestra, impacto y configuraciones de los parámetros de los ítemes. Los procedimientos bietápicos condujeron a resultados similares a los estándar en presencia de impacto. El método que generalmente arroja menores proporciones de falsos positivos es la PN; le siguen ambos tipos de clasificación las cuales, teniendo el mismo costo computacional que la prueba de Mantel-Haenszel, fallan mucho menos en la detección errónea del DIF.

Erroneous detection of Differential Item Functioning. A comparison of methods. Some methods to analyze Differential Item Functioning (DIF) are compared in terms of their false positive rates, namely the normal test (NT) for the difference of difficulty parameters, the Mantel-Haenszel chi-square test and the item classification in 'Type C' or 'Not Type C' according to the only magnitude of the MH D-DIF statistic (empirical classification) -the last two on both a single-stage and two-stage procedures- and according to ETS criterion. They were applied to data without DIF which were simulated under different design conditions with respect to sample size, impact and item parameter configurations. The results obtained from two-stage and single-stage procedures were similar when impact is present. The method that generally produces lower false positive rates is the NT one, followed by both types of classification which, holding the same computational cost as Mantel-Haenszel test, have a much lower failure rate in the erroneous detection of DIF.

Existen diversos procedimientos para el estudio del funcionamiento diferencial del ítem (DIF), los cuales pueden clasificarse entre los que aplican la Teoría de Respuesta al Ítem (TRI) y los llamados de tablas de contingencia (TC). Numerosos trabajos de simulación los comparan teniendo en cuenta la proporción de error de tipo I, otros también concluyen respecto de la potencia de los métodos. Están los que consideran que los grupos no difieren en habilidad (Kim, Cohen y Kim, 1994; Cohen, Kim y Wollack, 1996; Fidalgo, Mellenbergh y Muñiz, 1998; Fidalgo et al., 2000) y los que contemplan la presencia de impacto (Cohen y Kim, 1993; Fidalgo et al., 1999), en todos estos estudios ambos grupos tienen igual tamaño de muestra. La situación de desbalance entre los grupos cuando se estudia el DIF sobre la base de datos reales es frecuente (Gómez y Navas, 1998; Elosúa y López, 1999; Elosúa, López y Egaña, 2000; Lozzia, Galibert, Aguerri y Attorresi, 2003) y ha sido contemplada desde la simulación en menos ocasiones. Entre tales trabajos puede mencionarse a Donoghue, Holland y Thayer (1993) y Narayaman y Swaminathan (1994), quienes aplican el procedimiento de Mantel-Haenszel a ítemes dicotómicos, y a Kim y Cohen (1998), que utilizan el método de la razón de verosimilitud para estudiar el DIF de ítemes de respuesta graduada.

En el presente trabajo interesa comparar los resultados de diversos métodos de detección del DIF cuando no lo hay en ítemes dicotómicos. Los métodos utilizados son la prueba normal para la diferencia de los parámetros de dificultad y tres criterios de análisis del DIF encuadrados dentro de los métodos TC: la prueba χ^2 de Mantel-Haenszel, la clasificación del ítem en 'tipo C', o no, según el criterio utilizado en el Educational Testing Service (ETS), y según dicho criterio restringido al aspecto descriptivo que considera sólo si el valor absoluto del estadístico MH D-DIF es mayor o igual que 1.5. Ésta fue denominada clasificación empírica por resultar de la utilizada por el ETS prescindiendo de la parte inferencial.

Los resultados del efecto del tamaño de muestra, la habilidad de los sujetos y las características de los ítemes en la detección errónea del DIF, falsos positivos, al utilizar la clasificación empírica se encuentran en Aguerri, Zanelli, Galibert y Attorresi (2002) y al emplear la prueba normal para la diferencia de los parámetros de dificultad en Attorresi, Galibert, Zanelli, Lozzia y Aguerri (2003).

Fecha recepción: 18-6-03 • Fecha aceptación: 3-11-04
Correspondencia: María Ester Aguerri
Facultad de Psicología
Universidad de Buenos Aires
Agrelo 3555. 1224 (Argentina)
E-mail: maguerri@psi.uba.ar

Metodología

Diseño y simulación de los datos

Los factores considerados en este trabajo son: tamaño de muestra, distribución de la habilidad, discriminación del ítem y dificultad del ítem. El grupo de referencia (GR) es, en todos los casos, de tamaño 900 y pertenece a una población cuya habilidad se distribuye como una normal estándar. Los grupos focales (GF) resultan de combinar dos niveles para el tamaño de muestra (N_{GF}= 900, N_{GF}= 350) y tres niveles para la habilidad media de la población a la cual pertenece el GF ($\mu_{\theta GF}$= -1.5, $\mu_{\theta GF}$= 0, $\mu_{\theta GF}$ = 1.5) En todos los casos se considera que la habilidad se distribuye normalmente con desvío estándar igual a 1.

Los datos fueron simulados con el modelo logístico de tres parámetros mediante un programa especialmente confeccionado en SAS (Statistical Analysis System, 1989) con los mismos parámetros generadores en ambos grupos, esto es, sin DIF. Algunos trabajos de simulación buscan inspirarse en datos reales para que la simulación sea lo más parecida a la realidad. Sin embargo, en datos reales algunas combinaciones entre los parámetros de dificultad y de discriminación son poco frecuentes: no es fácil lograr un ítem muy difícil, o muy fácil, de alta discriminación. No obstante puede resultar interesante ver cómo se comporta un método en tales situaciones. Por ello se ha elegido un diseño con valores prefijados para los parámetros en las marcas de clase que representan a cada nivel. Se consideró una prueba de 20 ítemes. El parámetro de aciertos por azar de todos los ítemes se fijó en 0.25. El valor del parámetro de discriminación y el del parámetro de dificultad del ítem resultan de cruzar cuatro niveles para el parámetro de discriminación (0.4, 0.8, 1.2 y 1.6) con cinco niveles para el parámetro de dificultad (-2, -1, 0, 1 y 2).

Para cada una de las 2 × 3 × 4 × 5 combinaciones de los niveles de los factores se consideraron 50 repeticiones en las que se estudió la detección errónea del DIF.

Prueba normal para la diferencia de los parámetros de dificultad

La prueba normal para la diferencia de los parámetros de dificultad, propuesta por Wright, Mead y Draba (1976), contrasta las hipótesis H_0: $\Delta b = b_R - b_F = 0$, H_1: $\Delta b = b_R - b_F \neq 0$ donde b_R es el parámetro de dificultad para el ítem en el GR y b_F lo es en el GF. El estadístico de prueba se distribuye asintóticamente como una normal estándar, razón por la cual en este trabajo se menciona a esta prueba como prueba normal (PN). El análisis del DIF se efectuó con BILOG-MG™ (Zimowski, Muraki, Mislevy y Bock, 1996). Este programa utiliza el método de estimación de los parámetros de máxima verosimilitud marginal. El procedimiento elegido consistió en ajustar el modelo de tres parámetros a cada grupo considerando para cada ítem que el parámetro c, de aciertos por azar, es el mismo para los dos grupos, así como también es igual la potencia discriminatoria del ítem en los dos grupos, es decir: $c_R = c_F$ y $a_R = a_F$. El programa proporciona los valores de $\hat{b}$ ajustados para cada grupo y la diferencia de los mismos con su respectivo error estándar. Con estos resultados se estudió el DIF de los ítemes mediante la prueba normal, con un nivel de significación de 0.05.

Métodos del tipo TC

* Se aplicó la prueba χ^2 de Mantel-Haenszel (Mantel y Haenszel, 1959; Holland y Thayer, 1986, 1988) mediante el procedimiento estándar implementado a partir del Proc. Freq. de SAS (Statistical Analysis System, 1989) y mediante un procedimiento bietápico de depuración del criterio con el programa EZDIF (Waller, 1998) indicados como MH-1 y MH-2, respectivamente. En este trabajo se pretende evaluar la posible ventaja de los procedimientos bietápicos respecto de los procedimientos estándar; Fidalgo, Mellenbergh y Muñiz (1998, 1999), a partir de un diseño diferente, recomiendan la aplicación del procedimiento bietápico de Mantel-Haenszel tanto en cuanto al error de tipo I como a la potencia.

* Se registró si el ítem fue erróneamente clasificado 'tipo C' o no a partir del programa EZDIF (Waller, 1998). El Educational Testing Service utiliza una clasificación de los ítemes según su DIF sobre la base de la magnitud y significancia del estadístico MH D-DIF, definido por Holland y Thayer (1988), como ítemes tipo A: con DIF muy pequeño, B: con DIF intermedio, o C: con DIF grande (Zieky, 1993). Un ítem es clasificado 'tipo C' si el valor absoluto del estadístico MH D-DIF es mayor o igual que 1.5 y significativamente mayor que 1.

* Se consideró un procedimiento basado sólo en el aspecto descriptivo de la clasificación 'tipo C' que considera que el ítem exhibe DIF si el valor absoluto del estadístico MH D-DIF es mayor o igual que 1.5. Ésta fue denominada clasificación empírica por resultar del criterio utilizado por el ETS prescindiendo de la parte inferencial. Este recurso es de sencilla implementación pues basta con observar el valor absoluto del estadístico MH D-DIF; ha sido aplicado en el estudio del DIF en casos reales por Ferreres, González-Romá y Gómez (2002, p. 462), quienes señalan que: «Los ítems con magnitud (módulo del estadístico delta MH-D) mayor o igual que 1.5 son los que presentan problemas de funcionamiento diferencial». El programa MHDIF (Fidalgo, 1994) no ofrece la clasificación 'tipo C' pero sí informa sobre la magnitud del estadístico MH D-DIF. Tal valor también es informado por el programa EZDIF (Waller, 1998). El SAS (Statistical Analysis System, 1989), a este respecto, sólo brinda la estimación de la razón común de las posibilidades α_{MH} en la salida del PROC FREQ, dicha estimación convenientemente transformada permite obtener el valor del estadístico MH D-DIF. En este trabajo se aplicó la clasificación empírica de manera estándar (CE-1) y bietápica (CE-2), mediante la aplicación del SAS (Statistical Analysis System, 1989) y el programa EZDIF (Waller, 1998), respectivamente.

El estudio de la detección errónea del DIF según la clasificación en 'tipo C' o no del ítem apunta a evaluar la bondad de la clasificación empírica.

Resultados

La proporción de DIF erróneamente detectado con los métodos PN, MH-1, MH-2, la clasificación 'tipo C' , CE-1 y CE-2 según el tamaño de muestra del GF, el valor del parámetro de discriminación y de dificultad del ítem se presenta en la tabla 1 cuando el GF está en marcada desventaja en cuanto a la habilidad, en la tabla 2 para el caso en que los grupos no difieren en cuanto a la habilidad y en la tabla 3 cuando el GF presenta marcada ventaja en cuanto a la habilidad.

Proporción de detección errónea con la prueba normal para la diferencia de los parámetros de dificultad

En la tabla 1 puede apreciarse que la proporción de detección errónea al estudiar el DIF con la prueba normal toma valores superiores a 0.05 sólo cuando N_{GF}= 350 en ítemes muy difíciles (b= 2) de discriminación superior a 0.4 y difíciles de discriminación alta (b= 1 y a= 1.6). Este resultado podría atribuirse a la inexactitud de las estimaciones por ser reducida la cantidad de sujetos en los niveles más altos de habilidad, en Attorresi et al. (2003) se muestra que la recuperación de los parámetros en tales situaciones resultó pobre. Cuando los grupos no difieren en habilidad o el GF presenta marcada ventaja la proporción de detección errónea con PN toma valores que no superan 0.05.

La proporción de DIF erróneamente detectado con PN verifica la condición liberal de Bradley (1978), esto es, toma un valor comprendido entre 0.025 y 0.075, en el 15% de los ítemes en presencia de impacto y en un 5% cuando los grupos sólo difieren en tamaño de muestra. Cuando los grupos no difieren en tamaño ni en habilidad ningún ítem satisface la condición liberal de Bradley, pues en todos los casos el nivel de significación empírico no supera a 0.02.

El ajuste del modelo logístico de tres parámetros a los datos resultó adecuado sólo cuando los ítemes tienen baja discriminación (a= 0.4) o son muy difíciles (b= 2) con la excepción del caso en que el GF presenta marcada ventaja en cuanto a la habilidad y tiene tamaño 900. En muchos de los casos contemplados en este diseño se rechazó el ajuste en alta proporción, siendo menos rechazado cuando el GF es de tamaño 350.

Prueba χ^2 de Mantel-Haenszel

Según se aprecia en las tablas 1, 2 y 3 la proporción de detección errónea con esta prueba resulta muy influenciada por el tamaño de muestra y la presencia o no de impacto. El efecto es diferente según los niveles del parámetro de discriminación del ítem y el sentido del impacto. Por lo general se observa que a mayor tamaño de muestra del grupo focal le corresponde una mayor proporción de detección errónea. Por otra parte, puede apreciarse que la utilización del procedimiento bietápico condujo a resultados similares a los del procedimiento estándar en presencia de impacto. Cuando el GF está en desventaja el valor medio para MH-1 es 0.429 y para MH-2 es 0.4247, mientras que tales valores cuando el GF está en ventaja son, respectivamente, 0.2278 y 0.2033. Los valores más bajos de proporción de detección errónea corresponden a la situación en la que los grupos no difieren en habilidad (Tabla 2) con un valor promedio de 0.068 para MH-1 y 0.0455 para MH-2. En presencia de impacto contra GF (Tabla 1) se observan valores de proporción de DIF erróneamente detectado 'inflados' en ítemes de discriminación baja que no son fáciles (a= 0.4 y b≥0), valores que aumentan según aumenta la dificultad del ítem. La proporción de detección errónea está altamente 'inflada' en los ítemes de discriminación superior a 0.4, exceptuando a casi todos los ítemes de dificultad intermedia. Los resultados observados en cuanto a la incidencia de los parámetros del ítem y del tamaño de muestra en la prueba χ^2 de Mantel-Haenszel se corresponden con las tendencias registradas por Roussos y Stout (1996) en similar situación. Cuando el GF está en marcada ventaja en cuanto a la habilidad (Tabla 3) los ítemes de discriminación baja (a= 0.4) presentan la tendencia observada en la situación de impacto opuesto. Los ítemes de discriminación medio-baja (a= 0.8) presentan valores altos para la proporción de DIF erróneamente detectado con MH-1 y MH-2 cuando el parámetro de dificultad toma valores extremos (b= -2 o b= 2), particularmente cuando N_{GF}= 900. Mientras que cuando el parámetro de discriminación del ítem es medio-alto y alto (a≥1.2) son los ítemes de dificultad extrema los que presentan valores relativamente más bajos de detección errónea.

La proporción de DIF erróneamente detectado con la prueba χ^2 de Mantel-Haenszel satisfizo la condición liberal de Bradley (1978) sólo en el 15% de las 120 situaciones estudiadas cuando se aplicó MH-1. Al aplicar MH-2 la fracción de ítemes que verifica la condición mencionada aumentó al 25%. Los ítemes que verifican la condición liberal de Bradley en su mayoría son ítemes de

Tabla 1
Proporción de DIF erróneamente detectado según el método, el tamaño de muestra del GF y los parámetros del ítem cuando el GF está en marcada desventaja en cuanto a la habilidad

		a= 0.4					a= 0.8					a= 1.2					a= 1.6				
		b= -2	b= -1	b= 0	b= 1	b= 2	b= -2	b= -1	b= 0	b= 1	b= 2	b= -2	b= -1	b= 0	b= 1	b= 2	b= -2	b= -1	b= 0	b= 1	b= 2
PN	N_{GF}= 900	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.04	0.04	0.04	0.00	0.02	0.02	0.00	0.00
	N_{GF}= 350	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	**0.10**	0.02	0.02	0.00	0.02	**0.07**	0.02	0.04	0.04	**0.08**	**0.13**
MH-1	N_{GF}= 900	0.04	0.12	**0.20**	**0.36**	**0.64**	**0.56**	**0.44**	0.08	**0.36**	**0.60**	**0.98**	**0.90**	0.12	**0.30**	**0.79**	**1.00**	**1.00**	0.14	**0.39**	**0.84**
	N_{GF}= 350	0.10	0.04	**0.14**	**0.24**	**0.28**	**0.48**	**0.30**	0.06	**0.18**	**0.52**	**0.82**	**0.62**	0.04	**0.28**	**0.57**	**0.88**	**0.8**6	0.08	**0.29**	**0.53**
MH-2	N_{GF}= 900	0.04	0.06	**0.10**	**0.28**	**0.62**	**0.68**	**0.54**	0.14	**0.32**	**0.60**	**0.98**	**0.98**	0.18	**0.28**	**0.77**	**1.00**	**0.98**	0.28	**0.33**	**0.84**
	N_{GF}= 350	0.04	0.06	**0.16**	**0.14**	**0.22**	**0.50**	**0.34**	0.08	**0.18**	**0.50**	**0.80**	**0.62**	0.04	**0.22**	**0.54**	**0.86**	**0.84**	0.12	**0.29**	**0.45**
Tipo C	N_{GF}= 900	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	**0.46**	**0.16**	0.00	0.00	0.02	**0.66**	**0.28**	0.00	0.00	0.00
	N_{GF}= 350	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	**0.28**	**0.08**	0.00	0.00	0.02	**0.32**	**0.18**	0.00	0.00	0.00
CE-1	N_{GF}= 900	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	**0.62**	**0.10**	0.00	0.00	0.02	**0.82**	**0.30**	0.00	0.00	0.02
	N_{GF}= 350	0.00	0.00	0.00	0.00	0.02	0.06	0.00	0.00	0.02	0.06	**0.58**	**0.10**	0.02	0.00	0.04	**0.72**	**0.30**	0.00	0.00	0.09
CE-2	N_{GF}= 900	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	**0.72**	**0.18**	0.00	0.00	0.04	**0.90**	**0.54**	0.00	0.00	0.02
	N_{GF}= 350	0.00	0.00	0.00	0.00	0.00	**0.14**	0.02	0.00	0.00	0.06	**0.60**	**0.16**	0.00	0.00	0.09	**0.72**	**0.48**	0.00	0.00	0.09

discriminación baja y medio-baja. Es interesante destacar que la situación más favorable para el cumplimiento de tal condición se verifica cuando no hay presencia de impacto, con MH-2 un 50% de los ítemes si los grupos no difieren en tamaño de muestra y un 40% cuando sí difieren versus el 20% y 25%, respectivamente, con MH-1.

Clasificación 'tipo C'

En las tablas 1, 2 y 3 se evidencia el efecto de los parámetros del ítem según el sentido y grado del impacto. Cuando GR y GF no difieren en cuanto a la habilidad (Tabla 2) la proporción de clasificación 'tipo C' errónea es nula salvo cuando el ítem es de discriminación medio-alta y muy fácil (a= 1.2 y b= -2) y el GF es de tamaño 350 donde vale 0.02. En la situación en la que el GF está en marcada desventaja (Tabla 1) los ítemes de discriminación medio-alta y alta y dificultad inferior a la intermedia (a≥1.2 y b<0) son los que presentan una proporción de clasificación errónea notoriamente 'inflada'. Cuando el GF tiene marcada ventaja en cuanto a la habilidad y es de tamaño 900 (Tabla 3) son los ítemes fáciles de discriminación medio-alta y alta (b= -1 y a≥1.2) y los de dificultad intermedia y alta discriminación (a= 1.6 y b= 0) los clasificados 'tipo C' erróneamente en una mayor proporción. Cuando el GF presenta marcada desventaja la proporción máxima de clasificación 'tipo C' errónea es 0.66, con valor promedio de 0.082 para N_{GF}= 900 y de 0.045 para N_{GF}= 350, mientras que cuando GF presenta marcada ventaja tales valores son 0.18, 0.024 y 0.005, respectivamente.

Tabla 2
Proporción de DIF erróneamente detectado según el método, el tamaño de muestra y los parámetros del ítem cuando los grupos GR y GF no difieren en cuanto a la habilidad

		a= 0.4					a= 0.8					a= 1.2					a= 1.6				
		b= -2	b= -1	b= 0	b= 1	b= 2	b= -2	b= -1	b= 0	b= 1	b= 2	b= -2	b= -1	b= 0	b= 1	b= 2	b= -2	b= -1	b= 0	b= 1	b= 2
PN	NGF= 900	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	NGF= 350	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.02	0.00	0.02	0.04	0.00	0.00	0.00	0.00	0.02	0.00
MH-1	NGF= 900	0.02	0.02	0.08	0.10	0.10	0.10	0.04	0.12	0.08	0.10	0.12	0.10	0.02	0.10	0.12	0.04	0.10	0.06	0.02	0.08
	NGF= 350	0.08	0.12	0.06	0.00	0.08	0.06	0.00	0.06	0.10	0.06	0.08	0.12	0.02	0.06	0.02	0.12	0.02	0.12	0.00	0.02
MH-2	NGF= 900	0.02	0.02	0.04	0.06	0.04	0.06	0.00	0.12	0.08	0.06	0.06	0.10	0.04	0.04	0.12	0.00	0.04	0.08	0.00	0.04
	NGF= 350	0.06	0.06	0.04	0.00	0.02	0.02	0.00	0.06	0.08	0.06	0.06	0.08	0.02	0.06	0.02	0.06	0.00	0.10	0.00	0.00
Tipo C	NGF= 900	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	NGF= 350	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CE-1	NGF= 900	0.00	0.02	0.02	0.00	0.00	0.02	0.00	0.02	0.00	0.00	0.10	0.04	0.02	0.00	0.00	0.02	0.02	0.02	0.02	0.02
	NGF= 350	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	**0.22**	0.00	0.02	0.00	0.00
CE-2	NGF= 900	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	NGF= 350	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	**0.16**	0.00	0.00	0.00	0.00

Tabla 3
Proporción de DIF erróneamente detectado según el método, el tamaño de muestra del GF y los parámetros del ítem cuando el GF presenta marcada ventaja en cuanto a la habilidad

		a= 0.4					a= 0.8					a= 1.2					a= 1.6				
		b= -2	b= -1	b= 0	b= 1	b= 2	b= -2	b= -1	b= 0	b= 1	b= 2	b= -2	b= -1	b= 0	b= 1	b= 2	b= -2	b= -1	b= 0	b= 1	b= 2
PN	NGF= 900	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.02	0.00	0.02	0.04	0.04	0.00	0.00	0.00	0.04	0.02
	NGF= 350	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.04	0.04	0.04
MH-1	NGF= 900	0.06	**0.16**	**0.24**	**0.50**	**0.52**	**0.24**	0.10	0.04	0.06	**0.32**	0.07	**0.50**	**0.58**	**0.20**	**0.24**	0.00	**0.60**	**0.86**	**0.46**	**0.14**
	NGF= 350	0.08	0.08	**0.18**	**0.24**	**0.36**	0.02	0.06	0.06	0.00	**0.14**	0.00	**0.26**	**0.24**	**0.14**	**0.12**	0.00	**0.31**	**0.60**	**0.24**	**0.10**
MH-2	NGF= 900	0.08	**0.16**	**0.20**	**0.38**	**0.50**	**0.10**	0.04	0.08	0.04	**0.24**	0.04	**0.52**	**0.64**	**0.18**	**0.20**	0.02	**0.60**	**0.82**	**0.48**	**0.12**
	NGF= 350	0.06	0.06	**0.18**	**0.14**	**0.36**	0.00	0.08	0.06	0.00	**0.10**	0.00	**0.14**	**0.26**	**0.14**	0.06	0.00	**0.14**	**0.44**	**0.34**	0.08
Tipo C	NGF= 900	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.02	0.00	0.00	0.00	**0.16**	**0.18**	0.00	0.00
	NGF= 350	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.02	0.00	0.00	0.00	0.00	0.04	0.00	0.00
CE-1	NGF= 900	0.00	0.00	0.00	0.00	0.00	**0.20**	0.02	0.00	0.00	0.00	**0.44**	**0.40**	0.06	0.00	0.00	**0.48**	**0.64**	**0.28**	0.00	0.00
	NGF= 350	0.04	0.00	0.00	0.00	0.02	**0.30**	0.06	0.00	0.00	0.00	**0.57**	**0.42**	0.06	0.00	0.00	**0.15**	**0.65**	**0.28**	0.00	0.00
CE-2	NGF= 900	0.00	0.00	0.00	0.00	0.00	**0.30**	0.02	0.00	0.00	0.00	**0.61**	**0.50**	0.06	0.00	0.00	**0.60**	**0.68**	**0.34**	0.00	0.00
	NGF= 350	0.06	0.00	0.00	0.00	0.02	**0.40**	0.10	0.02	0.00	0.00	**0.50**	**0.48**	0.10	0.02	0.00	**0.55**	**0.61**	**0.38**	0.06	0.00

Clasificación empírica

La proporción de detección errónea con la clasificación empírica, tanto en la versión estándar como bietápica, parece ser sensible a la presencia o no de impacto y a una combinación particular de los parámetros de discriminación y dificultad del ítem. Son los ítemes muy fáciles de discriminación medio-alta y alta (b= -2 y a≥1.2) los que presentan valores de proporción de detección errónea 'inflada' en todas las situaciones, con valores máximos observados de 0.90, 0.61 y 0.22 cuando el GF está en marcada desventaja, en marcada ventaja o no difiere con GR en cuanto a la habilidad. En la situación en la que no hay impacto (Tabla 2) el método presenta los valores más altos de detección errónea en ítemes muy fáciles y discriminatorios (b= -2 y a= 1.6) cuando el grupo focal es el de menor tamaño, en una proporción 0.22 para CE-1 y 0.16 para CE-2, los respectivos valores medios son 0.0165 y 0.0085. En presencia de impacto la clasificación empírica también exhibe valores de proporción de detección errónea 'inflada', particularmente CE-2, en ítemes fáciles de discriminación medio-alta y alta (b=-1 y a≥1.2). Cuando el GF tiene marcada ventaja (Tabla 3) se agregan los ítemes de dificultad intermedia y discriminación alta (b=0 y a=1.6) y los muy fáciles con discriminación medio-baja (b= -2 y a= 0.8). En el caso que el GF está en desventaja el promedio para CE-1 es 0.0978 y 0.1214 para CE-2, mientras que tales valores cuando el GF está en ventaja son, respectivamente, 0.1265 y 0.1602.

Se observa correspondencia entre los resultados de la clasificación empírica y la clasificación 'tipo C' en presencia de impacto; la excepción la constituyen los ítemes muy fáciles cuando el GF está en ventaja en cuanto a la habilidad, situación en que la clasificación 'tipo C' no falla. Roussos y Stout (1996), para el caso en que la diferencia entre las medias de las poblaciones es 1 y el parámetro de aciertos por azar es 0.2, encontraron que ítemes sin DIF, fáciles y de alta discriminación (a= 2.5 y b= -1.5) presentan valores del estadístico MH D-DIF en módulo superiores a 1.5.

Según los resultados obtenidos en este estudio puede afirmarse que conviene basarse en la magnitud del DIF (clasificación empírica y clasificación 'tipo C') antes de descartar a un ítem a partir del procedimiento de Mantel-Haenszel acorde a lo considerado por Fidalgo y Ferreres (2002). Esto es especialmente recomendable en presencia de impacto cuando los ítemes presentan dificultad superior a la intermedia.

Limitaciones del presente estudio

Este trabajo se restringe al estudio de la detección errónea del DIF, falsos positivos y no aborda el estudio de las identificaciones correctas de situaciones en las que el ítem presenta DIF. Cohen y Kim (1993) estudian la potencia de tres métodos, de los cuales el más comparable al de la prueba normal es el referido al área con signo Z(*ESA)* de Raju, aunque lo aplican al modelo de dos parámetros. En cuanto a la prueba de Mantel-Haenszel puede recurrirse a los estudios sobre la potencia presentados en Narayaman y Swaminathan (1994) y Fidalgo et al. (1999). Estos estudios confirman, aunque sobre diferentes diseños, que la potencia aumenta con el tamaño de muestra.

Conclusiones

Los factores contemplados en este diseño: tamaño de muestra del GF, presencia o no de impacto, y los parámetros del ítem resultaron influyentes en la proporción de detección errónea con diferente intensidad y sentido según el método empleado.

La proporción de DIF erróneo correspondiente a la prueba normal tiende en general a estar muy por debajo del nivel de significación esperado; excepto en cuatro casos muy puntuales cuando el grupo focal es minoritario y está en desventaja en cuanto a la habilidad. Estos ítemes corresponden a las siguientes configuraciones de los parámetros: a≥0.8 y b= 2, y a= 1.2 y b= 1, situación atribuible a la inexactitud de las estimaciones.

Respecto de los métodos TC se aprecia la marcada incidencia del impacto y de los parámetros del ítem en la proporción de detección errónea. El tamaño de muestra del GF influyó en la proporción de DIF erróneamente detectado con la prueba χ^2 de Mantel-Haenszel en presencia de impacto, pues se observa que a mayor tamaño de muestra le corresponde mayor valor de dicha proporción. Estas afirmaciones coinciden con resultados obtenidos para la prueba χ^2 de Mantel-Haenszel por Roussos y Stout (1996) para el caso en que la diferencia entre las medias de las poblaciones es 1 y el parámetro de aciertos por azar es 0.2.

Como es de esperar, los métodos tienden a cometer error de tipo I en menor proporción cuando no hay impacto; habiendo rendido resultados óptimos la clasificación 'tipo C' como también su versión restringida —clasificación empírica— exceptuando los ítemes muy fáciles y discriminatorios cuando el grupo focal es el de menor tamaño. En presencia de impacto contra el grupo focal ambos métodos funcionan satisfactoriamente para casi todos los ítemes y de manera muy similar, lo que habilitaría a usar la clasificación empírica, más sencilla. En este caso de impacto hay determinadas configuraciones de los parámetros de los ítemes que tienden a arrojar proporciones de detección errónea más altas: son ítemes fáciles o muy fáciles y discriminatorios (b<0 y a ≥ 1.2), en los que todos los métodos TC fallan, o bien ítemes muy difíciles y extremadamente discriminatorios (b= 2 y a= 1.6) cuando el grupo focal es minoritario, donde fallan todos los métodos excepto la clasificación 'tipo C'. Cuando el grupo favorecido es el focal, los métodos TC también fallan para los ítemes discriminatorios como en el caso anterior, pero ahora en ítemes fáciles o de dificultad media (b= -1, b= 0). Según los resultados obtenidos en este trabajo puede afirmarse que conviene basarse en la magnitud del DIF (clasificación empírica y clasificación 'tipo C') antes de descartar a un ítem a partir del procedimiento de Mantel-Haenszel acorde a lo considerado por Fidalgo y Ferreres (2002). Esto es especialmente recomendable en presencia de impacto cuando los ítemes presentan dificultad superior a la intermedia.

Los procedimientos bietápicos condujeron a resultados similares a los procedimientos estándar, particularmente en situación de impacto.

De este trabajo surge que el método para detectar DIF más recomendable, en cuanto al riesgo de cometer error de tipo I, es, en condiciones generales, la prueba normal para la diferencia de los parámetros de dificultad. Si no se dispone del bagaje conceptual y del soporte computacional que involucra el estudio del DIF con los métodos de la Teoría de Respuesta al Ítem, puede recurrirse a la clasificación 'tipo C' o a la clasificación empírica, particularmente cuando los grupos no difieren en cuanto a la habilidad. Ambos tipos de clasificación tienen el mismo costo computacional que la prueba de Mantel-Haenszel, pero fallan mucho menos en la detección errónea del DIF.

Agradecimientos

Esta investigación fue realizada con subsidios de la Universidad de Buenos Aires (UBACyT P054, P605, P020 y P027), del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET PIP Nº 2426) y de la Agencia Nacional de Promoción Científica y Tecnológica (PICT Nº 4704).

Referencias

Aguerri, M.E., Zanelli, M.L., Galibert, M.S. y Attorresi, H.F. (2002). Evaluación de un método empírico para detectar el funcionamiento diferencial del ítem. *Interdisciplinaria, Revista de Psicología y Ciencias Afines, Journal of Psychology and Related Sciencies, 19*(2), pp. 185-213.

Attorresi, H.F., Galibert, M.S., Zanelli, M.L., Lozzia, G.S. y Aguerri, M.E. (2003). Error de tipo I en el análisis del Funcionamiento Diferencial del Ítem basado en la diferencia de los parámetros de dificultad. *Psicológica, 24*(2), 289-306.

Bradley, J.V. (1978). Robustness? *The British Journal of Mathematical & Statistical Psychology, 31,* 144-152.

Cohen, A.S. y Kim, S.-H. (1993). A comparison of Lord's χ^2 and Raju's Areas Measures in detection of DIF. *Applied Psychological Measurement, 17*(1), 39-52.

Cohen, A.S., Kim, S.-H. y Wollack, A. (1996). An Investigation of the Likelihood Ratio Test for Detecting Differential Item Functioning. *Applied Psychological* Measurement, *20*, 15-26.

Donoghue, J.R., Holland, W.P. y Thayer, D.T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. En P.W. Holland y H. Wainer (Eds.): *Differential Item Functioning* (pp. 137-166) Hillsdale, NJ: Erlbaum.

Elosúa, P. y López, A. (1999). Funcionamiento diferencial de los ítems y sesgo en la adaptación de dos pruebas verbales. *Psicológica, 20,* 23-40.

Elosúa, P., López, A. y Egaña, J. (2000). Idioma de aplicación y rendimiento en una prueba de comprensión verbal. *Psicothema, 12*(2), 201-206.

Ferreres, D., González-Romá, V. y Gómez, J. (2002). Funcionamiento diferencial de los ítems en una situación de contacto de lenguas. *Psicothema, 14*(2), 483-490.

Fidalgo, A.M. (1994). MHDIF: a computer program for detecting uniform and nonuniform differential item functioning with the Mantel-Haenszel procedure. *Applied Psychological Measurement, 18,* 300.

Fidalgo, A. y Ferreres, D. (2002). Supuestos y consideraciones en los estudios empíricos sobre el funcionamiento diferencial de los ítems. *Psicothema, 14*(2), 491-496.

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*(1), 209-218.

Fidalgo, A.M., Mellenbergh, G.J. y Muñiz, J. (1999). Aplicación en una etapa, dos etapas e iterativamente de los estadísticos de Mantel-Haenszel. *Psicológica, 20,* 227-242.

Fidalgo, A.M., Mellenbergh, G.J. y Muñiz, J. (2000). Effects of amount of DIF, test length and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online, 5*, 3. http://www.mpr-online.de.

Gómez, J. y Navas, M.J. (1998). Impacto y funcionamiento diferencial de los ítems respecto al género en una prueba de aptitud numérica. *Psicothema, 10*(3), 685-696.

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

Holland, P.W. y Thayer, D.T. (1988) Differential item functioning and the Mantel-Haenszel procedure. En H. Wainer y H.I.Braun (Eds.): *Test Validity* (pp. 129 -145) Hillsdale, NJ: Lawrence Erlbaum.

Kim, S.-H., Cohen, A.S. y Kim, H.-O. (1994). An investigation of Lord's procedure for detection of differential item functioning. *Applied Psychological Measurement, 18*(3), 217-228.

Kim, S.-H. y Cohen, A.S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*(4), 345-355.

Lozzia, G.S., Galibert, M.S., Aguerri, M.E. y Attorresi, H.F. (2003). Construcción de un banco de ítem de razonamiento verbal. *Interdisciplinaria, Revista de Psicología y Ciencias Afines, Psychology and Related Sciencies.* Aceptado para su publicación 01-10-03.

Mantel N. y Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Narayaman, P. y Swaminathan, H. (1994) Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*(4), 315-328.

Roussos, L. y Stout, W. (1996) Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*(2), 215-230.

SAS Institute Inc. (1989). SAS/STAT® *User's Guide*. Version 6, Fourth Edition,volume 1, Cary, N.C.: SAS Institute Inc., 943 pp.

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

Wright, B.D., Mead, R. y Draba R. (1976). *Detecting and correcting test item bias with a logistic response model* (Research Memorandum N.º 22). Chicago: The University of Chicago, Department of Education, Statistical Laboratory.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. En P.W. Holland y H. Wainer (Eds.): *Differential Item Functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

Zimowski, M., Muraki, E., Mislevy, R. y Bock, R. (1996). *BILOG-MG™: Multiple-Group IRT Analysis and Test Maintenance for Binary Items* [Computer program] Scientific Software International, Inc.