SECCIÓN METODOLÓGICA

Psicológica (2003), 24, 289-306.

Error de Tipo I en el análisis del Funcionamiento Diferencial del Ítem basado en la diferencia de los parámetros de dificultad

Horacio F. Attorresi*[1], María Silvia Galibert*, Marta L. Zanelli**, Gabriela S. Lozzia* y María Ester Aguerri*

* Universidad de Buenos Aires. ** Instituto Nacional de Tecnología Agropecuaria (Argentina)

Se estudia mediante simulación el Error de Tipo I cometido en el análisis del Funcionamiento Diferencial del Ítem (DIF) cuando se aplica la prueba normal para la diferencia de los parámetros de dificultad. En el diseño se consideraron diversas situaciones en cuanto al tamaño de muestra para los grupos focal y de referencia (iguales o diferentes), a la distribución de la habilidad en las respectivas poblaciones (iguales o en dos situaciones de discrepancia) y en cuanto a los parámetros de los ítemes que se eligieron combinando distintos niveles de discriminación y de dificultad. La proporción de DIF erróneamente detectado se mantuvo por debajo del nivel de significación de 0.05 en un 96% de los casos; las situaciones en las que lo superó corresponde a ítemes difíciles respondidos por sujetos de bajos niveles de habilidad. La prueba normal para la diferencia de los parámetros de dificultad es de sencilla implementación y permite mantener controlado el riesgo de descartar erróneamente ítemes por su DIF aunque el Error de Tipo II podría ser mayor que el esperado.

En el estudio del funcionamiento diferencial del ítem (DIF) se comparan las respuestas de sujetos de distintos grupos a un ítem. Se detecta DIF cuando sujetos de un mismo nivel de habilidad tienen distinta probabilidad de contestar correctamente el ítem, según el grupo al que pertenezcan. Hambleton y Swaminathan (1985), Thissen, Steinberg, y Wainer (1993), Camilli y Shepard (1994), entre otros autores, presentan los procedimientos estadísticos utilizados para el análisis del DIF que se basan en la aplicación de los modelos de la Teoría de Respuesta al Ítem (TRI). Están los que

[1] Esta investigación fue realizada con los siguientes subsidios: Universidad de Buenos Aires (UBACYT P054/00), del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET PIP 2426/00) y de la Agencia Nacional de Promoción Científica y Tecnológica (PICT 4704/98). Sede del Proyecto: Instituto de Investigaciones. Facultad de Psicología de la Universidad de Buenos Aires. Correspondencia: Horacio Félix Attorresi. Rivera Indarte 132, 1er. Piso, Dpto. A, (1406) Buenos Aires, Argentina. E-mail: hattorre@psi.uba.ar

comparan los parámetros del ítem estimados en los dos grupos, los que comparan las curvas características del ítem estimadas para cada grupo mediante la medición del área comprendida entre las mismas y los que usan la comparación del ajuste de los modelos. Al primer grupo corresponden la prueba normal para los parámetros de dificultad propuesta por Wright, Mead, y Draba (1976) y la prueba χ^2 de Lord (1977, 1980). En la segunda categoría se encuadran los métodos de comparación de áreas como el de Raju (1988, 1990) y en la tercera se encuentra la prueba de la razón de verosimilitud, Thissen et al. (1988, 1993).

Cohen y Kim (1993) presentan una comparación del procedimiento χ^2 de Lord con las pruebas Z(*ESA*) y Z(*H*) de medidas del área de Raju con signo y sin signo respectivamente. Los datos simulados siguen el modelo de dos parámetros para ítemes dicotómicos. Estudian el efecto del tamaño de muestra, de la longitud del test y de la presencia o no de impacto, esto es que los grupos pertenezcan a poblaciones que difieren en cuanto a la habilidad o no. Sólo estudian el DIF en los casos en que ambos grupos tienen igual tamaño de muestra. El diseño considera tests con tres porcentajes de ítemes con DIF: 0%, 10% y 20%. Para los tres estadísticos encuentran un número de falsos positivos (detecciones erróneas del DIF) inferior al esperado según el nivel nominal cuando hay un 0% de ítemes con DIF en el test. Concluyen que si bien la diferencia entre los tres estadísticos es leve, el procedimiento χ^2 de Lord aventaja a los otros dos.

Kim, Cohen, y Kim (1994) estudian el efecto de dos algoritmos de estimación de los parámetros del ítem: máxima verosimilitud marginal y bayesiana marginal sobre la razón de Error de Tipo I al aplicar el procedimiento χ^2 de Lord. Los datos fueron simulados según un diseño que considera dos tamaños de muestra (1000 y 250) que se mantienen iguales entre los grupos de individuos, cuya habilidad se simuló a partir de una distribución normal estándar. Ajustan tres modelos de la TRI: dos parámetros (2PLM), tres parámetros (3PLM) y tres parámetros con parámetro de aciertos por azar fijo (3PLM-c). Concluyen que la prueba χ^2 de Lord no provee un buen control del Error de Tipo I cuando se ajusta el 3PLM; por el contrario la proporción de Error de Tipo I se mantiene más baja que el nivel nominal al ajustar el 2PLM o el 3PLM-c. Asimismo observan que la proporción de Error de Tipo I es inferior a la que se obtendría suponiendo conocida la habilidad –caso en que dicha proporción se mantiene más próxima al nivel nominal- y sugieren que esto se debe a una sobreestimación de los errores estándar.

Cohen, Kim, y Wollack (1996) investigan la proporción de Error de Tipo I en la detección del DIF al aplicar la prueba de la razón de máxima verosimilitud. Este estudio lo hacen con las mismas condiciones en el diseño y sobre los mismos datos simulados que utilizaron Kim et al. (1994) con el fin de comparar los resultados de la prueba de la razón de máxima verosimilitud con la prueba χ^2 de Lord. Es decir que en todos los casos considerados en este trabajo los grupos pertenecen a una población cuya habilidad se distribuye como una normal estándar y no se efectúan estudios del DIF cuando los tamaños de muestra de los grupos son diferentes. La

prueba de la razón de verosimiltud presenta una proporción de DIF erróneamente detectado más próximo al nivel de significación que la prueba χ^2 de Lord en todos los casos.

Kim y Cohen (1998) examinan la proporción de Error de Tipo I en la detección del DIF al aplicar la prueba de la razón de verosimilitud a datos simulados según el modelo de respuesta graduada. En el diseño cruzan dos tamaños de muestra y dos situaciones en cuanto a la habilidad, con y sin impacto. La proporción de Error de Tipo I analizada según seis niveles de significación: 0.0005, 0.001, 0.005, 0.01, 0.05 y 0.1, resulta próxima al valor esperado en todas las condiciones.

Camilli y Shepard (1994) distinguen el concepto *de funcionamiento diferencial del ítem* del concepto de *sesgo del ítem*. Mientras el primero es puramente estadístico, el segundo considera las causas de tal funcionamiento diferencial. Se hará referencia al sesgo de los ítemes sólo cuando se hayan dado explicaciones debidamente fundadas para el funcionamiento diferencial. Así, el análisis del funcionamiento diferencial de los ítemes (DIF) puede ser útil no sólo para la creación de instrumentos de medición invariantes entre poblaciones sino también para detectar diferencias entre grupos cuyas interpretaciones podrían generar hipótesis de interés. Pero antes de aventurarse a una interpretación convendría tener cierta seguridad de que no se está en presencia de un falso DIF; por lo que resulta de interés investigar si determinadas condiciones -como los parámetros de los ítemes, tamaños de muestra y diferencia en la habilidad de los grupos- pueden ser factores de riesgo en cuanto a aumentar la proporción de DIF erróneo. Estas consideraciones conducen, por tanto, a la necesidad de analizar la proporción de Error de Tipo I del DIF, dejando el estudio del Error de Tipo II para una etapa posterior de la investigación.

Siguiendo la línea de investigación de los autores antes mencionados el propósito de este trabajo es analizar la proporción de Error de Tipo I en la detección del DIF pero cuando el procedimiento estadístico utilizado es la prueba normal para la diferencia de los parámetros de dificultad. A diferencia de dichos autores, que centran su interés en la comparación de los métodos, en el presente trabajo interesa también sacar conclusiones en cuanto a la incidencia de los parámetros del ítem sobre dicha proporción; por lo que se lleva a cabo un diseño que considera todas las combinaciones de los diversos niveles para los parámetros de dificultad y de discriminación. Se estudian diferentes casos dependiendo de que los grupos difieran o no en cuanto al tamaño de muestra y en cuanto a la habilidad media. Interesó estudiar la incidencia que podría tener una diferencia destacada entre los niveles medios de habilidad de los grupos sospechando que esto podría aumentar la proporción de Error de Tipo I y conducir a interpretaciones falaces en términos de sesgo. Los resultados que se presentan en este trabajo son parte de un estudio más extenso sobre el DIF que se encuentra en Aguerri (2000).

MÉTODO

Diseño y simulación de los datos.

En la literatura sobre el DIF un grupo suele identificarse como grupo de Referencia (GR) y el otro como grupo Focal (GF). El grupo de Referencia es de tamaño 900 y pertenece a una población cuya habilidad se distribuye como una normal estándar. Los grupos Focales combinan dos posibles tamaños de muestras: 900 y 350, y tres posibles medias para la habilidad: en igualdad de condiciones respecto del GR, y en situación de discrepancia, tanto que aventaje al GR como que esté en desventaja. En la Tabla 1 se muestran los seis tipos de GF diferenciados por el tamaño de muestra y la media de la distribución de la habilidad en la población a la cual pertenecen. Los tamaños de muestra elegidos, 900 y 350, están en el orden de otros trabajos de simulación como Kim y Cohen (1998) que utilizaron 1000 y 300, y Fidalgo, Mellenbergh, y Muñiz (1999) que eligieron 1000 y 200. En el diseño se consideró la posibilidad de desbalance entre GR y GF, situación que es habitual en la práctica (Nandakumar, 1993; Zwick, Thayer, y Lewis, 1999; Elosua y López, 1999; Galibert, 2000; Bielinski, Thurlow, Ysseldyke, Freidebach, y Freidebach, 2001).

Tabla 1. Identificación de los Grupos Focales, según el tamaño de muestra (n_{GF}) y la media de la habilidad de la población a la cual pertenecen ($\mu_{\theta GF}$).

$\mu_{\theta GF}$ \ n_{GF}	900	350
-1.5	Grupo Focal 1	Grupo Focal 4
0	Grupo Focal 2	Grupo Focal 5
1.5	Grupo Focal 3	Grupo Focal 6

Los datos fueron simulados con el modelo logístico de tres parámetros mediante un programa especialmente confeccionado en SAS (Statistical Analysis System, 1989). Este programa requiere que se especifiquen los parámetros de los ítemes para los que se simulan las respuestas de los sujetos. En principio a cada sujeto se le asigna aleatoriamente un nivel de habilidad según la población a la que pertenece. Posteriormente se calcula la probabilidad de que conteste correctamente el ítem mediante el modelo logístico de tres parámetros:

$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-1,7a_i(\theta - b_i)}}, \theta$$

y se genera para cada sujeto y para cada ítem un número al azar entre 0 y 1. Si el número generado es menor o igual que la probabilidad calculada se considera que el sujeto acierta el ítem y se le asigna 1; de lo contrario se le asigna 0. Así se obtiene una matriz de 1 y 0, con tantas columnas como ítemes y tantas filas como sujetos.

Se consideraron cuatro niveles para el parámetro de discriminación: Bajo, Medio-Bajo, Medio-Alto y Alto, y cinco niveles para el parámetro de dificultad: Muy Fácil, Fácil, Medio, Difícil y Muy Difícil. El parámetro de acierto por azar se fijó en 0.25 dado que es frecuente en la práctica psicométrica utilizar de ítemes del tipo de elección múltiple con cuatro alternativas. Los valores de los parámetros de discriminación y dificultad de los ítemes se muestran en la Tabla 2.

Se simularon 50 repeticiones para cada una de las 2x3x4x5 combinaciones de los niveles de los factores: Tamaño de muestra, Distribución de la habilidad, Discriminación del ítem y Dificultad del ítem.

Tabla 2. Identificación de los ítemes según la combinación de los valores de los parámetros de dificultad (b) y de discriminación (a).

a \ b	-2	-1	0	1	2
0.4	. Item 1	Item 2	Item 3	Item 4	Item 5
0.8	Item 6	Item 7	Item 8	Item 9	Item 10
1.2	Item 11	Item 12	Item 13	Item 14	Item 15
1.6	Item 16	Item 17	Item 18	Item 19	Item 20

Prueba normal para la diferencia de los parámetros de dificultad.

Se contrastan 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, Z, se obtiene dividiendo la diferencia $\Delta\hat{b} = \hat{b}_R - \hat{b}_F$ por su error estándar. Es decir $Z = \Delta\hat{b} / s_{\Delta\hat{b}}$ con $s_{\Delta\hat{b}} = \sqrt{s^2_{\hat{b}_R} + s^2_{\hat{b}_F}}$ donde, bajo H_0, Z se distribuye asintóticamente como una normal estándar.

Esta prueba se deriva de la aplicación de técnicas multivariadas que permiten la comparación de los vectores que contienen los parámetros del ítem. Hambleton y Swaminathan (1985) muestran cómo se reduce la expresión general al caso en el que se ajusta el modelo de Rasch y llegan al estadístico presentado por Wright et al. (1976). Se requiere como supuesto necesario que los estimadores de los parámetros de dificultad en ambos grupos sean independientes y normalmente distribuidos. Raju (1990) deduce, entre otros, el mismo estadístico a partir del método de las áreas para el

modelo logístico de tres parámetros; es el correspondiente al área con signo. Se necesita suponer que el parámetro de aciertos por azar es igual entre los grupos. 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 descripto en Baker (1992). El procedimiento elegido consiste 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 $a_R = a_F$. Luego la comparación de los parámetros de los ítemes se reduce a la comparación de los parámetros de dificultad; esto es el DIF uniforme. La prueba normal para la diferencia de los parámetros de dificultad fue aplicada a datos reales, entre otros autores, por Draba (1977), Schulz (1990), Galibert (2000) y Bielinski et al. (2001).

Implementación computacional.

El programa provee estimaciones diferentes del parámetro de dificultad para cada grupo, luego efectúa un reescalamiento de manera que la media de los parámetros de dificultad para el GR sea cero; muestra los valores de β 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, para lo cual se efectuaron cálculos en Statistix® for Windows (1996).

RESULTADOS

Problemas de convergencia.

En algunas de las corridas computacionales requeridas por el diseño se presentaron problemas de convergencia en el proceso de estimación de los parámetros. Como lo hicieran Cohen et al. (1996, pp.20) ante dichos problemas se optó por excluir de la corrida al ítem que obstaculiza la convergencia y modificar las condiciones 'por default', en este trabajo se aumentó el criterio de convergencia. En la Tabla 3 se registra para cada ítem el número de corridas entre las 50 en las que se alcanzó la convergencia según las características de sus parámetros y del grupo Focal. Cuando el grupo de Referencia y el grupo Focal no difieren en cuanto a la habilidad media todas las corridas resultan válidas. Las corridas se reducen para ítemes difíciles cuando el grupo Focal está en desventaja, y para los ítemes fáciles cuando el grupo Focal está en ventaja. Claramente esto se debe a la falta de variabilidad en las respuestas que tienden a ser incorrectas en el primer caso y correctas en el segundo. La simetría entre estas dos situaciones no es completa, sin embargo, dado que el parámetro de aciertos por azar otorga alguna probabilidad de respuesta correcta al grupo desaventajado. Por tanto, mientras que se conservó alguna variabilidad en las respuestas de este grupo, no fue así en varios de los patrones simulados para el grupo aventajado, donde en

muchos casos las respuestas de todos los sujetos a ítemes fáciles resultaron correctas. Hubo, por tanto, una mayor reducción de corridas en este caso, que llegan a ser sólo 30 y 20 para los ítemes 11 y 16 respectivamente en el GF6.

Tabla 3. Cantidad de repeticiones válidas para cada ítem según el tamaño de muestra del Grupo Focal (n_{GF}), la media de la habilidad de la población a la cual dicho grupo pertenece ($\mu_{\theta GF}$) y el valor del parámetro de discriminación (a) y de dificultad (b). Se destacan en negrita los valores menores de 50.

		n_{GF}=900					n_{GF}=350				
		b=-2	b=-1	b=0	b=1	b=2	b=-2	b=-1	b=0	b=1	b=2
$\mu_{\theta GF}$=-1.5	a = 0.4	50	50	50	50	50	50	50	50	50	50
	a = 0.8	50	50	50	50	50	50	50	50	50	**48**
	a =1.2	50	50	50	50	**48**	50	50	50	50	**46**
	a = 1.6	50	50	50	**49**	**49**	50	50	50	**49**	**47**
$\mu_{\theta GF}$=0	a = 0.4	50	50	50	50	50	50	50	50	50	50
	a = 0.8	50	50	50	50	50	50	50	50	50	50
	a = 1.2	50	50	50	50	50	50	50	50	50	50
	a = 1.6	50	50	50	50	50	50	50	50	50	50
$\mu_{\theta GF}$=1.5	a = 0.4	50	50	50	50	50	50	50	50	50	50
	a = 0.8	50	50	50	50	50	50	50	50	50	50
	a = 1.2	**46**	50	50	50	50	**30**	50	50	50	50
	a = 1.6	**42**	50	50	50	50	**20**	**49**	50	50	50

Recuperación de los parámetros.

La evaluación de la recuperación de los parámetros se efectuó mediante la raíz del error cuadrático medio (RECM) para los parámetros del ítem y la correlación entre las estimaciones de los parámetros con sus respectivos valores generadores. La correlación para las estimaciones del parámetro de aciertos por azar no se realizó puesto que el valor generador en todos los casos fue 0.25. Los resultados se exhiben en las Tablas 4, 5, 6, 7 y 8. Los valores más bajos de la RECM corresponden al parámetro de aciertos por azar (Tabla 6). Para el parámetro de discriminación se observa en la Tabla 4 que la RECM aumenta con la dificultad del ítem, particularmente cuando la potencia discriminatoria es alta. Los valores mayores corresponden a la situación en la que el grupo Focal está en marcada desventaja en cuanto a la habilidad. En la Tabla 5 puede apreciarse que los valores más bajos de la RECM para el parámetro de dificultad corresponden a ítemes de dificultad intermedia y los valores más altos a ítemes difíciles, salvo que el grupo

presente una marcada ventaja en cuanto a la habilidad. En dicha situación los valores altos de la RECM se encuentran en los ítemes más fáciles. En la Tabla 7 se observa que las correlaciones entre las estimaciones del parámetro de discriminación y los valores generadores son altas salvo cuando los ítemes son difíciles, particularmente cuando el grupo Focal presenta una marcada desventaja en cuanto a la habilidad. La correlación entre las estimaciones del parámetro de dificultad y los valores generadores resulta alta en todos los casos, según puede observarse en la Tabla 8. El tamaño de muestra afectó sólo levemente la recuperación de los parámetros mientras que la dificultad del ítem y la presencia de impacto condujeron a estimaciones más imprecisas.

Proporción de DIF erróneo.

En la Tabla 9 puede observarse que la proporción de DIF erróneo detectado con la prueba normal se mantiene dentro del nivel de significación elegido en 116 de las 120 combinaciones de los niveles de los factores. La proporción de DIF erróneo que excede al 0.05 se observa en cuatro ítemes, cuando el grupo focal es minoritario y desaventajado en cuanto a la habilidad. La configuración de los parámetros de dichos ítemes es: muy alta dificultad (b=2) y discriminación medio-baja (a=0.8), muy alta dificultad (b=2) y discriminación medio-alta (a=1.2), alta dificultad (b=1) y alta discriminación (a=1.6) y muy alta dificultad (b=2) y alta discriminación (a=1.6). En las Tablas 5, 6 y 8 puede observarse que en dichas situaciones la recuperación de los parámetros resultó pobre, posiblemente por ser reducida la cantidad de sujetos en los niveles más altos de habilidad. Al examinar por qué esta situación no se da en el caso simétrico de ítemes fáciles contestados por un grupo focal aventajado, se observó que el error estándar del estimador del parámetro de discriminación aumenta, en promedio, con el nivel de discriminación del ítem conjuntamente con su dificultad, más allá de las características de los grupos (Figuras 1 y 2). Asimismo, en cuanto a las estimaciones del parámetro de acierto por azar, las Figuras 3 y 4 muestran una ligera tendencia decreciente al aumentar conjuntamente la dificultad y la discriminación del ítem. Dichas estimaciones son muy similares entre los grupos cuando los ítemes son difíciles, y discrepan más entre sí cuando los ítemes son fáciles, caso donde los errores estándar son mayores. Esta incidencia de la configuración de parámetros de dificultad y discriminación sobre las estimaciones y sus errores estándar parece explicar, en parte, que no se den resultados simétricos en cuanto a la detección errónea del DIF cuando se considera un grupo aventajado contestando a un ítem fácil. Por otra parte, la proporción de DIF erróneamente detectado en estos casos fue calculada sobre una base menor según se señalara oportunamente en los problemas de convergencia, lo que hace más inestable dicha proporción. Más aún, dado que la cantidad de veces que se detecta DIF al 5% en una cierta cantidad “n” de corridas puede suponerse una variable Binomial de parámetros n y p=0.05, la probabilidad de no detectar DIF por encima del nivel de significación es 0.74 para el caso del ítem 16 contestado por el GF6 contra 0.58 en la situación simétrica del ítem 20 al ser contestado por el GF4.

Tabla 4. Raíz del error cuadrático medio del parámetro de discriminación del ítem según la media de la habilidad de la población a la cual pertenece el GF ($\mu_{\theta GF}$) y el valor del parámetro de discriminación (a) y de dificultad (b) y el tamaño de muestra de dicho grupo (n_{GF}). Se destacan en negrita los casos correspondientes a la detección de DIF erróneo.

			b=-2	b=-1	b=0	b=1	b=2
$\mu_{\theta GF}$=-1.5	a=0.4	n_{GF}=900	0.059	0.085	0.085	0.109	0.136
		n_{GF}=350	0.063	0.085	0.090	0.108	0.126
	a=0.8	n_{GF}=900	0.088	0.121	0.136	0.135	0.244
		n_{GF}=350	0.092	0.120	0.144	0.147	**0.246**
	a=1.2	n_{GF}=900	0.171	0.190	0.235	0.226	0.482
		n_{GF}=350	0.174	0.189	0.240	0.250	**0.514**
	a=1.6	n_{GF}=900	0.178	0.231	0.229	0.370	0.869
		n_{GF}=350	0.197	0.204	0.244	**0.412**	**0.857**
$\mu_{\theta GF}$=0	a=0.4	n_{GF}=900	0.053	0.061	0.092	0.109	0.114
		n_{GF}=350	0.060	0.079	0.102	0.135	0.163
	a=0.8	n_{GF}=900	0.084	0.096	0.120	0.135	0.215
		n_{GF}=350	0.107	0.107	0.130	0.144	0.218
	a=1.2	n_{GF}=900	0.180	0.171	0.191	0.173	0.369
		n_{GF}=350	0.194	0.229	0.271	0.208	0.403
	a=1.6	n_{GF}=900	0.210	0.182	0.200	0.308	0.624
		n_{GF}=350	0.258	0.196	0.259	0.344	0.747
$\mu_{\theta GF}$=1.5	a=0.4	n_{GF}=900	0.058	0.081	0.081	0.073	0.073
		n_{GF}=350	0.062	0.069	0.095	0.088	0.122
	a=0.8	n_{GF}=900	0.097	0.102	0.094	0.110	0.134
		n_{GF}=350	0.112	0.095	0.113	0.121	0.176
	a=1.2	n_{GF}=900	0.213	0.247	0.264	0.145	0.275
		n_{GF}=350	0.204	0.200	0.218	0.191	0.315
	a=1.6	n_{GF}=900	0.299	0.209	0.222	0.213	0.358
		n_{GF}=350	0.308	0.214	0.270	0.261	0.448

Tabla 5. Raíz del error cuadrático medio del parámetro de dificultad del ítem en el Grupo de Referencia (RECMbGR) y en el Grupo Focal (RECMbGF), según la habilidad media de la población a la cual pertenece el GF ($\mu_{\theta GF}$), el tamaño de muestra de dicho grupo (n_{GF}), el valor del parámetro de discriminación (a) y de dificultad (b). Se destacan en negrita los casos correspondientes a la detección de DIF erróneo.

				b=-2	b=-1	b=0	b=1	b=2
$\mu_{\theta GF}$=-1.5	a=0.4	RECMbGR	n_{GF}=900	0.310	0.236	0.177	0.229	0.316
			n_{GF}=350	0.303	0.279	0.221	0.249	0.378
		RECMbGF	n_{GF}=900	0.316	0.261	0.276	0.302	0.554
			n_{GF}=350	0.398	0.426	0.461	0.459	0.641
	a=0.8	RECMbGR	n_{GF}=900	0.192	0.139	0.147	0.137	0.377
			n_{GF}=350	0.214	0.153	0.154	0.169	**0.334**
		RECMbGF	n_{GF}=900	0.184	0.153	0.176	0.290	0.648
			n_{GF}=350	0.258	0.219	0.256	0.369	**0.595**
	a=1.2	RECMbGR	n_{GF}=900	0.186	0.116	0.089	0.100	0.617
			n_{GF}=350	0.194	0.127	0.100	0.120	**0.506**
		RECMbGF	n_{GF}=900	0.155	0.140	0.142	0.238	0.716
			n_{GF}=350	0.210	0.175	0.175	0.329	**0.554**
	a=1.6	RECMbGR	n_{GF}=900	0.142	0.091	0.026	0.105	1.031
			n_{GF}=350	0.150	0.098	0.102	**0.102**	**0.771**
		RECMbGF	n_{GF}=900	0.158	0.127	0.107	0.360	0.924
			n_{GF}=350	0.192	0.160	0.182	**0.385**	**0.699**
$\mu_{\theta GF}$=0	a=0.4	RECMbGR	n_{GF}=900	0.285	0.281	0.272	0.252	0.275
			n_{GF}=350	0.310	0.254	0.254	0.216	0.337
		RECMbGF	n_{GF}=900	0.304	0.316	0.264	0.230	0.318
			n_{GF}=350	0.372	0.333	0.342	0.312	0.422
	a=0.8	RECMbGR	n_{GF}=900	0.191	0.148	0.148	0.148	0.241
			n_{GF}=350	0.227	0.135	0.142	0.172	0.245
		RECMbGF	n_{GF}=900	0.173	0.165	0.136	0.185	0.295
			n_{GF}=350	0.250	0.174	0.173	0.171	0.353
	a=1.2	RECMbGR	n_{GF}=900	0.175	0.122	0.100	0.104	0.302
			n_{GF}=350	0.217	0.124	0.105	0.121	0.300
		RECMbGF	n_{GF}=900	0.198	0.134	0.103	0.092	0.285
			n_{GF}=350	0.269	0.187	0.127	0.130	0.382
	a=1.6	RECMbGR	n_{GF}=900	0.134	0.082	0.100	0.094	0.496
			n_{GF}=350	0.175	0.094	0.106	0.107	0.486
		RECMbGF	n_{GF}=900	0.164	0.115	0.094	0.097	0.449
			n_{GF}=350	0.266	0.132	0.138	0.157	0.411
$\mu_{\theta GF}$=1.5	a=0.4	RECMbGR	n_{GF}=900	0.376	0.340	0.282	0.255	0.332
			n_{GF}=350	0.530	0.397	0.385	0.304	0.221
		RECMbGF	n_{GF}=900	0.471	0.459	0.354	0.241	0.214
			n_{GF}=350	0.339	0.274	0.262	0.249	0.331

$\mu_{\theta GF}$=1.5	a=0.8	RECMbGR	n_{GF}=900	0.226	0.166	0.131	0.168 0.214
			n_{GF}=350	0.413	0.295	0.189	0.173 0.159
		RECMbGF	n_{GF}=900	0.348	0.264	0.137	0.129 0.134
			n_{GF}=350	0.239	0.130	0.132	0.161 0.236
	a=1.2	RECMbGR	n_{GF}=900	0.283	0.145	0.098	0.117 0.215
			n_{GF}=350	0.473	0.296	0.157	0.152 0.140
		RECMbGF	n_{GF}=900	0.506	0.233	0.148	0.106 0.117
			n_{GF}=350	0.270	0.104	0.083	0.115 0.250
	a=1.6	RECMbGR	n_{GF}=900	0.225	0.101	0.100	0.110 0.227
			n_{GF}=350	0.314	0.224	0.149	0.143 0.144
		RECMbGF	n_{GF}=900	0.364	0.207	0.134	0.097 0.104
				0.240	0.102	0.109	0.106 0.260

Tabla 6. Raíz del error cuadrático medio del parámetro de aciertos por azar del ítem según la media de la habilidad de la población a la cual pertenece el GF ($\mu_{\theta GF}$), el tamaño de muestra de dicho grupo (n_{GF}), el valor del parámetro de discriminación (a) y de dificultad del ítem (b).

			b=-2	b=-1	b=0	b=1	b=2
$\mu_{\theta GF}$=-1.5	a=0.4	n_{GF}=900	0.042	0.039	0.042	0.042	0.033
		n_{GF}=350	0.039	0.049	0.048	0.041	0.044
	a=0.8	n_{GF}=900	0.039	0.042	0.033	0.028	0.028
		n_{GF}=350	0.040	0.049	0.044	0.037	0.041
	a=1.2	n_{GF}=900	0.036	0.042	0.028	0.026	0.035
		n_{GF}=350	0.044	0.047	0.032	0.032	0.041
	a=1.6	n_{GF}=900	0.044	0.030	0.017	0.020	0.042
		n_{GF}=350	0.039	0.044	0.033	0.026	0.037
$\mu_{\theta GF}$=0	a=0.4	n_{GF}=900	0.050	0.058	0.060	0.058	0.037
		n_{GF}=350	0.046	0.049	0.058	0.056	0.035
	a=0.8	n_{GF}=900	0.024	0.054	0.045	0.039	0.037
		n_{GF}=350	0.026	0.046	0.045	0.037	0.037
	a=1.2	n_{GF}=900	0.020	0.045	0.042	0.020	0.026
		n_{GF}=350	0.022	0.053	0.041	0.032	0.037
	a=1.6	n_{GF}=900	0.022	0.042	0.037	0.024	0.022
		n_{GF}=350	0.073	0.051	0.048	0.026	0.030
$\mu_{\theta GF}$=1.5	a=0.4	n_{GF}=900	0.062	0.070	0.064	0.051	0.035
		n_{GF}=350	0.050	0.057	0.059	0.050	0.035
	a=0.8	n_{GF}=900	0.039	0.057	0.040	0.039	0.032
		n_{GF}=350	0.028	0.041	0.041	0.036	0.035
	a=1.2	n_{GF}=900	0.022	0.058	0.039	0.026	0.026
		n_{GF}=350	0.014	0.044	0.033	0.030	0.030
	a=1.6	n_{GF}=900	0.020	0.045	0.046	0.022	0.020
		n_{GF}=350	0.014	0.039	0.047	0.026	0.024

Tabla 7. Correlación entre las estimaciones del parámetro de discriminación y los valores generadores según el parámetro de dificultad del ítem (b), el tamaño de muestra del Grupo Focal (n_{GF}) y la media de la habilidad de la población a la cual dicho grupo pertenece ($\mu_{\theta GF}$). Se destacan en negrita los casos correspondientes a la detección de DIF erróneo.

		b=-2	b=-1	b=0	b=1	b=2
$\mu_{\theta GF}$=-1.5	n_{GF}=900	0.9646	0.9416	0.9220	0.8966	0.4861
	n_{GF}=350	0.9564	0.9454	0.9113	**0.8753**	**0.5547**
$\mu_{\theta GF}$=0	n_{GF}=900	0.9512	0.9604	0.9363	0.9087	0.7408
	n_{GF}=350	0.9219	0.9377	0.8974	0.8804	0.6132
$\mu_{\theta GF}$=1.5	n_{GF}=900	0.9330	0.9341	0.9323	0.9483	0.8936
	n_{GF}=350	0.9273	0.9364	0.9184	0.9240	0.8269

Tabla 8. Correlación entre las estimaciones del parámetro de dificultad para el grupo de Referencia y para el grupo Focal y los respectivos valores generadores según el parámetro de discriminación del ítem (a), el tamaño de muestra del Grupo Focal (n_{GF}) y la media de la habilidad de la población a la cual dicho grupo pertenece ($\mu_{\theta GF}$). Se destacan en negrita los casos correspondientes a la detección de DIF erróneo.

			a=0.4	a=0.8	a=1.2	a=1.6
$\mu_{\theta GF}$=-1.5	n_{GF}=900	GR	0.986	0.990	0.982	0.963
		GF	0.970	0.974	0.966	0.947
	n_{GF}=350	GR	0.982	0.990	**0.986**	**0.977**
		GF	0.943	0.967	**0.969**	**0.952**
$\mu_{\theta GF}$=0	n_{GF}=900	GR	0.985	0.992	0.992	0.987
		GF	0.983	0.991	0.993	0.988
	n_{GF}=350	GR	0.984	0.992	0.991	0.990
		GF	0.970	0.987	0.987	0.988
$\mu_{\theta GF}$=1.5	n_{GF}=900	GR	0.981	0.992	0.992	0.993
		GF	0.977	0.988	0.985	0.991
	n_{GF}=350	GR	0.983	0.992	0.992	0.994
		GF	0.971	0.983	0.983	0.989

Tabla 9. Proporción de DIF erróneo detectado con la prueba normal para la diferencia de los parámetros de dificultad indicado según la media de la habilidad de la población a la cual dicho grupo pertenece ($\mu_{\theta GF}$), el tamaño de muestra del Grupo Focal (n_{GF}), el valor del parámetro de discriminación (a) y de dificultad del ítem (b). Se destacan en negrita los casos correspondientes a la detección de DIF erróneo.

			b=-2	b=-1	b=0	b=1	b=2
$\mu_{\theta GF}$=-1.5	a=0.4	n_{GF}=900	0.000	0.000	0.000	0.000	0.020
		n_{GF}=350	0.000	0.000	0.000	0.000	0.000
	a=0.8	n_{GF}=900	0.000	0.000	0.020	0.000	0.000
		n_{GF}=350	0.000	0.000	0.000	0.020	**0.104**
	a=1.2	n_{GF}=900	0.000	0.020	0.040	0.040	0.042
		n_{GF}=350	0.020	0.020	0.000	0.020	**0.065**
	a=1.6	n_{GF}=900	0.000	0.020	0.020	0.000	0.000
		n_{GF}=350	0.020	0.040	0.040	**0.082**	**0.128**
$\mu_{\theta GF}$=0	a=0.4	n_{GF}=900	0.000	0.000	0.000	0.000	0.000
		n_{GF}=350	0.000	0.000	0.000	0.000	0.000
	a=0.8	n_{GF}=900	0.000	0.000	0.000	0.020	0.020
		n_{GF}=350	0.020	0.000	0.000	0.020	0.000
	a=1.2	n_{GF}=900	0.000	0.000	0.000	0.020	0.000
		n_{GF}=350	0.020	0.000	0.020	0.040	0.000
	a=1.6	n_{GF}=900	0.000	0.000	0.000	0.000	0.000
		n_{GF}=350	0.000	0.000	0.000	0.020	0.000
$\mu_{\theta GF}$=1.5	a=0.4	n_{GF}=900	0.000	0.000	0.000	0.000	0.000
		n_{GF}=350	0.000	0.000	0.000	0.000	0.020
	a=0.8	n_{GF}=900	0.000	0.000	0.000	0.020	0.000
		n_{GF}=350	0.000	0.000	0.000	0.000	0.020
	a=1.2	n_{GF}=900	0.022	0.000	0.020	0.040	0.040
		n_{GF}=350	0.000	0.000	0.000	0.020	0.000
	a=1.6	n_{GF}=900	0.000	0.000	0.000	0.040	0.020
		n_{GF}=350	0.000	0.000	0.040	0.040	0.040

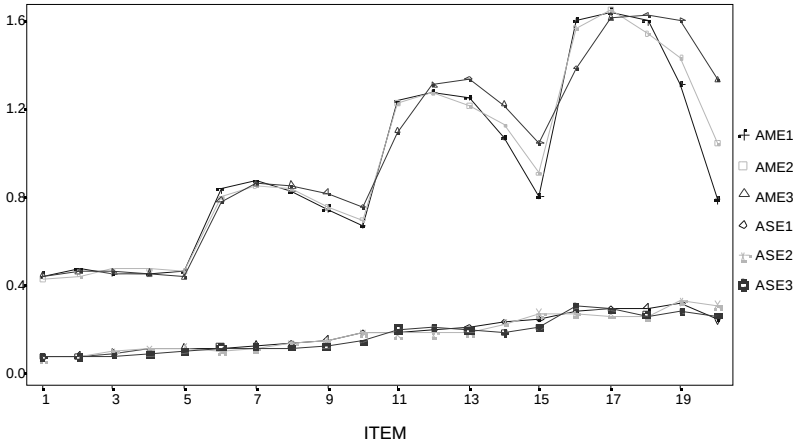

Figura 1 Promedio del parámetro de discriminación estimado (AME) y de su respectivo error estándar (ASE) en las repeticiones realizadas para cada item, para GF1,GF2 y GF3.

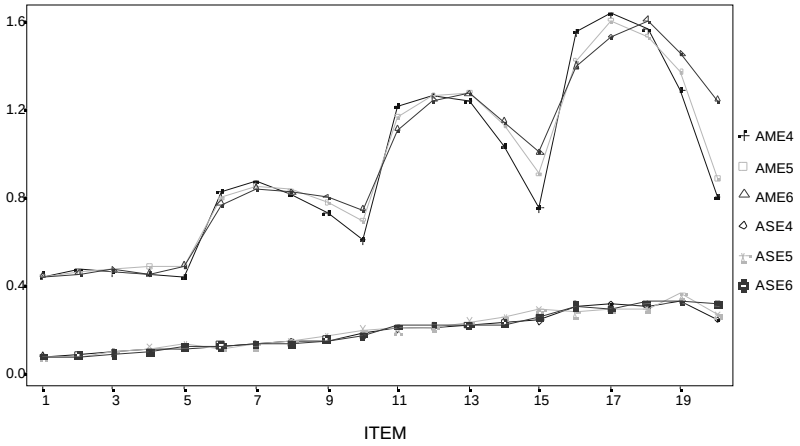

Figura 2 Promedio del parámetro de discriminación estimado (AME) y de su respectivo error estándar (ASE) en las repeticiones realizadas para cada item, para GF4, GF5 y GF6.

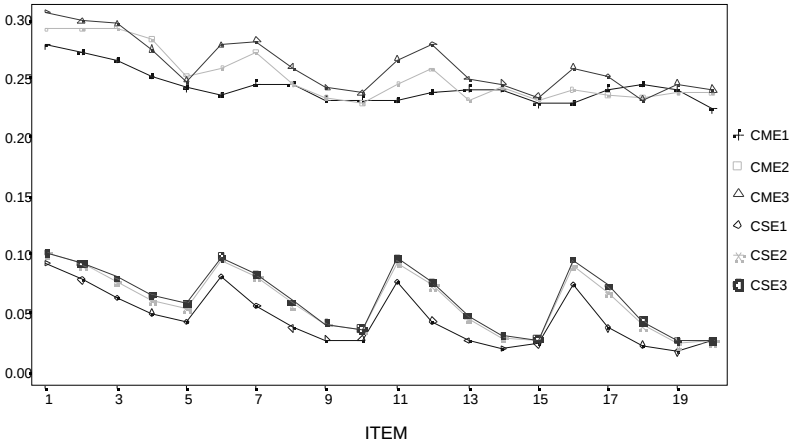

Figura 3 Promedio del parámetro de aciertos por azar estimado (CME) y de su respectivo error estándar (CSE) en las repeticiones realizadas para cada item, para GF1, GF2 y GF3.

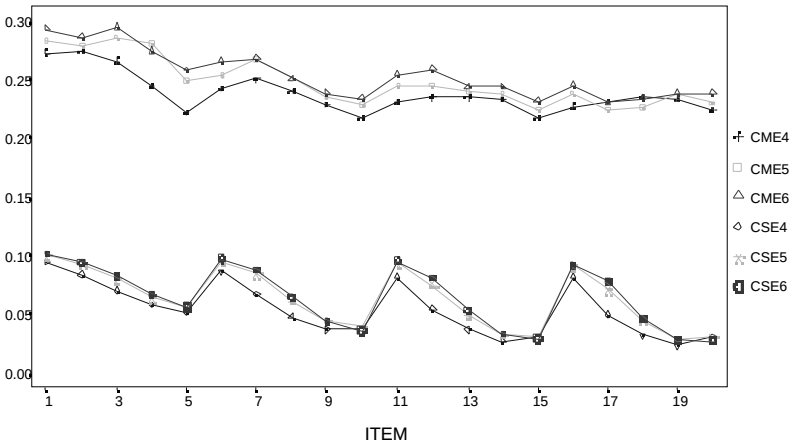

Figura 4 Promedio del parámetro de aciertos por azar estimado (CME) y de su respectivo error estándar (CSE) en las repeticiones realizadas para cada item, para GF4, GF5 y GF6.

Limitaciones del presente estudio.

Este trabajo se restringe al estudio del Error de Tipo I en la detección del DIF con la prueba normal para la diferencia de los parámetros de dificultad y no aborda el estudio del Error de Tipo II. 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. Para este método hallaron que un test de 20 ítemes presenta número promedio de 0,4 falsos negativos (no detección de DIF) a lo largo de cinco repeticiones si el test contiene un 10% de ítemes con DIF y un promedio de 3 falsos negativos si contiene un 20% de ítemes con DIF. Los grupos considerados no difieren en cuanto a la habilidad media y son ambos de 500 sujetos. Cuando los grupos difieren en la habilidad media los promedios de falsos negativos son 1.4 y 3.6 para un 10% y 20% de ítemes con DIF respectivamente. Todos los resultados precedentes corresponden a un nivel de significación del 5%. Como es de esperar, confirman que la potencia aumenta con el tamaño de muestra y el nivel de significación.

CONCLUSIONES

La proporción de DIF erróneamente detectado por la prueba normal para la diferencia de los parámetros de dificultad no superó al nivel de significación del 0.05 en el 96% de los casos. Como se señalara en la introducción, proporciones de DIF erróneo inferiores al nivel nominal también fueron halladas por Cohen y Kim (1994). El hecho de que las proporciones encontradas sean por lo general inferiores al valor nominal establecido puede atribuirse a una sobreestimación de los errores estándar dado que las estimaciones se llevan a cabo sin conocer los parámetros de habilidad; así lo observaron Kim et al. (1994) al ajustar los modelos 2PLM y 3PLM-c.

Los pocos casos en los que la proporción de DIF erróneamente detectado superó al nivel de significación se presentaron cuando ítemes difíciles y discriminatorios fueron respondidos por sujetos pertenecientes a un grupo focal de reducido tamaño y en marcada desventaja en cuanto a la habilidad (GF4). La situación simétrica corresponde a ítemes fáciles respondidos por sujetos aventajados (GF6). Esta simetría se reflejó parcialmente en los resultados relacionados con los problemas de convergencia pero no en cuanto a la recuperación de los parámetros ni a la detección errónea del DIF. La presencia de ítemes con valores extremos en los parámetros de dificultad y de discriminación puede traer problemas en la convergencia cuando se desea calibrar un conjunto de ítemes a partir de las respuestas de un grupo posicionado en el extremo opuesto a la dificultad del ítem en la escala de habilidad, problema que se presenta con más frecuencia conforme la muestra es de menor tamaño y en una dirección: respuestas a ítemes fáciles de sujetos de un grupo aventajado. La falta de simetría con respecto a la recuperación de los parámetros y a la proporción de DIF erróneo en las dos situaciones mencionadas posiblemente se explique porque no es simétrica la incidencia que tienen los diferentes niveles de dificultad y discriminación de los ítemes sobre las estimaciones. En efecto, hay una

aparente relación creciente entre los niveles de dificultad y de discriminación de los ítemes con el error estándar al estimar el parámetro de discriminación que sería de interés indagar en futuras investigaciones.

La prueba normal para la diferencia de los parámetros de dificultad es de sencilla implementación a partir de los resultados obtenidos en BILOG-MG™ y permite mantener controlado el riesgo de descartar erróneamente ítemes por su DIF. Sin embargo ha de utilizarse teniendo en cuenta que el nivel de significación real puede ser bastante inferior al nominal, por lo que el riesgo de cometer Error de Tipo II podría ser mayor que el esperado.

ABSTRACT

Type I Error in the Differential Item Functioning analysis based on the difficulty parameters difference. The Type I Error committed in the Differential Item Functioning (DIF) analysis when the Normal test is used for the difference of difficulty parameters is studied through simulation. In the design of this study different situations were considered with respect to: a) sample sizes of the focal and reference groups (equal or different); b) the ability distribution in the respective populations (equal or under two discrepancy situations) and c) the selection of different level combinations of the discrimination and difficulty item parameters. The proportion of DIF erroneously detected was kept under the 0.05 significance level in 96% of the cases, and those cases in which that significance level was exceeded are attributed to the lack of precision in the estimates due to insufficient number of observations for some ability levels. Therefore, the Normal test for the difference of difficulty parameters provides a high commendable DIF detection method from the point of view of the risk of indicating DIF when it does not exist.

REFERENCIAS

Aguerri, M. E. (2000). *Un estudio de simulación acerca del error de tipo I en la detección del funcionamiento diferencial del ítem.* Tesis de Magister Scientiae en Biometría. Universidad de Buenos Aires. Inédito.

Baker, F. (1992). *Item Response Theory. Parameter estimation techniques.* New York: Marcel Dekker.

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., y Freidebach, M. (2001). *Read-aloud accomodations: Effects on multiple-choice reading and math items* (Technical Report 31). Minneapolis, MN: University of Minesota, National Center of Outcomes. Retrieved (27/07/02), from World Wibe Web: http: \\education.umn.edu\NCEO\OnlinePubs\Technical31.htm

Camilli, G. y Shepard, L. (1994). *Methods for Identifying Biased Test Item*. Thousand Oaks, CA: Sage Publications, Inc..

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, J. A. (1996). An investigation of the likelihood ratio test for detecting differential item functioning. *Applied Psychological Measurement, 20,* 15-26.

Draba R. (1977). *The identification and interpretation of item bias* (Research Memorandum No. 25). Chicago: The University of Chicago, Dept. of Education, Statistical Lab.

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

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.

Galibert, M.S. (2000). *Modelización psicométrica de un test de razonamiento verbal en los marcos de la Teoría Clásica de Tests y de la Teoría de Respuesta al Ítem*. Tesis de Magister Scientiae en Biometría. Universidad de Buenos Aires. Inédito.

Hambleton, R. K. y Swaminthan, H. (1985). *Item Response Theory: Principles and Appplications.* Boston: Kluwer-Nijhoff.

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,* 345-356.

Lord , F.M. (1977). An study of item bias using item characteristic curve theory. En Y.H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19-29). Amsterdam, The Netherlands: Swets y Zeitlinger.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*, Hillsdale: Lawrence Erlbaum.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy- Stout's test for DIF. *Journal of Educational Measurement, 30,* 4, 293-311.

Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495-502.

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

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

Schulz, E. M. (1990). DIF detection: Rasch vs. Mantel-Haenszel. *Rasch Measurement Transactions, 4,* 2, 107.

Statistix® for Windows (1996). *User's Manual. Analytical Software*. Tallahassee, FL.

Thissen, D., Steinberg, L., y Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines, In H. Wainer y H.I. Braun (Eds.) *Test validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., y Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland y H. Wainer (Eds.) *Differential Item Functioning.* Hillsdale, NJ: Erlbaum.

Wright, B. D., Mead, R., y Draba R. (1976*). Detecting and correcting test item bias with a logistic response model* (Research Memorandum No. 22). Chicago: The University of Chicago, Department of Education, Statistical Laboratory.

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

Zwick, R., Thayer, D. y Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36,* 1, 1-28.

(Manuscrito recibido: 21/11/01; aceptado: 7/10/02)