Anuario de Psicología
1995, n.° *64*

Modelos lineales logarítmicos y funcionamiento diferencial de los ítems*

Ángel M. Fidalgo
M.ª Dolores Paz
Universidad de Oviedo

Se han utilizado modelos loglineales para la evaluación del comportamiento diferencial de los ítems (DIF). Para comprobar la efectividad de esta metodología se realizó un estudio de simulación en el que se manipuló el tamaño de la muestra, la cantidad de DIF y el nivel de significación elegido. Los resultados obtenidos muestran un alto porcentaje de ítems con DIF correctamente identificados, pero también de falsos positivos. Se sugiere una explicación de estos resultados en términos de la dimensionalidad del test.
Palabras clave: *Modelos loglineales, funcionamiento diferencial de los ítems, sesgo de los ítems, multidimensionalidad.*

Loglinear models were used to assess the effectiveness of the differential item functioning (DIF). A simulation study was performed in which sample size, amount of bias and significance levels were manipulated. High rates of correctly identified items with DIF were obtained, although there were also high numbers of false positives. The results are discussed in terms of test dimensionality.
Key words: *Loglinear models, Differential item functioning, Item bias, Multidimensionality.*

El uso de tests estandarizados en la evaluación de una gran variedad de variables psicológicas presenta como ventajas *a priori* la economía y facilidad de aplicación, su objetividad, y el supuesto de que, si se cumplen los

Dirección de los autores: Departamento de Filosofía y Psicología, Facultad de Psicología, Aniceto Sela, s/n. 33005 Oviedo.

*La realización del presente trabajo ha sido posible gracias a la Beca de Formación de Personal Investigador concedida al primer autor por la Dirección General de Investigación Científica y Técnica para llevar a cabo el proyecto de investigación titulado «Técnicas estadísticas para la evaluación del sesgo de los ítems».
Los autores agradecen al profesor Gideon J. Mellenbergh sus valiosos comentarios y sugerencias sobre el artículo.

requisitos psicométricos de fiabilidad y validez, proporcionarán medidas con poco error y válidas. Un test válido dará lugar a medidas idénticas para sujetos o grupos con iguales valores en la variable medida. Por el contrario, si la puntuación obtenida es función no sólo del nivel que los sujetos tienen en la variable medida, sino también de otras características irrelevantes como su pertenencia a diferentes grupos étnicos, culturales etc..., entonces hablamos de funcionamiento diferencial del test (DTF), y como tal constituye una causa clara de invalidez. Similarmente, con la expresión funcionamiento diferencial de los ítems (DIF) señalamos aquellos ítems cuya probabilidad de acertarlos difiere, a igual nivel en la variable medida, entre distintos subgrupos de una población dada. Cabría además distinguir entre DIF uniforme y no uniforme (Mellenbergh, 1982). El DIF uniforme se produce cuando la probabilidad de contestar correctamente a un ítem es mayor para un grupo que para otro a través de todos los niveles de habilidad. Por ejemplo, en un ítem que funcione diferencialmente en contra de las mujeres éstas tendrán una probabilidad de acertarlo menor, para todos los niveles de la variable medida, que los hombres con igual nivel de habilidad que ellas. El DIF no uniforme se produce cuando la diferencia en la probabilidad de responder correctamente a un ítem entre dos grupos no es la misma en todos los niveles de habilidad. En este caso no cabría hablar de DIF contra un grupo ya que, para determinados niveles de la habilidad, la probabilidad de acertar el ítem, a igual nivel en la variable medida, es mayor para un grupo, en tanto en los otros niveles es mayor para el otro.

Aunque el hecho de que un ítem funcione diferencialmente implica necesariamente una diferencia entre grupos, no de toda diferencia grupal en la ejecución de un ítem se sigue la presencia de DIF. Hay que distinguir entre el término DIF y el término impacto *(impact)*. Supongamos que los hombres tienen una mayor capacidad espacial que las mujeres. Si esto es cierto, los hombres, por término medio, obtendrán puntuaciones superiores en los tests de aptitud espacial, y también tendrán una probabilidad de acertar los ítems que componen los tests mayor que las mujeres. Sin embargo, los tests sólo muestran diferencias reales en la habilidad medida, estas diferencias se denominan impacto. Más formalmente, impacto, tomando la definición de Ackerman (1992), es una diferencia entre grupos en el desempeño en un ítem causada por una diferencia real en la variable medida.

Dadas las implicaciones de carácter ético, social y jurídico que conlleva la utilización de tests que pueden infravalorar sistemáticamente las capacidades de ciertos grupos en función de su cultura, etnia, sexo o cualquiera otra característica diferenciadora, no es extraño que últimamente una de las áreas más fecundas de la literatura psicométrica sea la dedicada a procurar procedimientos para detectar y eliminar aquellos ítems que presenten un funcionamiento diferencial. Entre los métodos propuestos para valorar el DIF cabría distinguir aquellos, formulados primeramente, que no controlan las diferencias en el nivel de habilidad entre grupos, como el análisis de varianza (Jensen, 1980; Osterlind, 1983) y el método delta (Angoff, 1982; Angoff y Ford, 1973; Angoff y Sharon, 1974), y aquellos más recientes que sí la controlan, estableciendo las comparaciones oportunas para iguales niveles de habilidad, tales

como chi-cuadrado (Camilli, 1979; Marascuilo, 1981; Marascuilo y Slaugther, 1981), Mantel-Haenszel (Holland y Thayer, 1986; 1988), modelos loglineales, logit y de clase latente (Kelderman, 1989; Kelderman y Macready, 1990; Kok, Mellenbergh y van der Flier, 1985; Mellenbergh, 1982; van der Flier, Mellenbergh, Adèr y Wijn, 1984; Westers y Kelderman, 1991), regresión logística (Spray y Carlson, 1986; Swaminathan y Rogers, 1990) y los métodos basados en la teoría de respuesta a los ítems (TRI) (Hambleton y Rogers, 1989; Kim y Cohen, 1991; Lord, 1980; Linn, Levine, Hastings y Wardrop, 1981; Muñiz, 1990; Raju, 1988, 1990; Rogers y Hambleton, 1989). En las primeras técnicas se puede producir confusión entre DIF e impacto, y llegar a concluir que ítems que reflejan la diferencia en la distribución de la variable medida entre grupos presentan DIF. Dentro del segundo grupo, las técnicas preferidas, y las más utilizadas, son los métodos basados en la TRI y el procedimiento Mantel-Haenszel (MH) propuesto por Holland y Thayer (1988). Uno de los inconvenientes de los métodos TRI es que son computacionalmente costosos además de requerir grandes tamaños de muestra (Hoover y Kolen, 1984). El procedimiento MH es particularmente atractivo por su fácil cálculo y el hecho de proporcionar una cuantificación del DIF presente en los ítems. Diversos estudios han encontrado, sin embargo, que no es un procedimiento eficaz cuando los ítems presentan DIF no uniforme (Hambleton y Rogers, 1989; Hills, 1989; Swaminathan y Rogers, 1990). Otra de las técnicas que sí controlan el nivel de competencia de los sujetos, como se ha señalado, son los modelos lineales logarítmicos o loglineales. Algunos autores han propuesto este tipo de técnicas para detectar el DIF, que permitirían, mediante la inclusión o no de determinados términos en los modelos, formular diferentes hipótesis sobre el tipo de DIF o las características de la distribución de las puntuaciones de los sujetos en el test (Green, Crone y Folk, 1989; Mellenbergh, 1982; Kelderman y Macready, 1990). Dentro de este marco, el objetivo del presente estudio es evaluar la eficacia de los modelos lineales logarítmicos en la detección de ítems que presenten DIF uniforme bajo diversas condiciones (diferente cantidad de DIF, diferentes tamaños de muestra y diferentes niveles de significación). Para ello se han utilizando datos simulados.

Método

Variables

Las variables manipuladas en esta investigación han sido:

a) Cantidad de DIF, definida como la diferencia entre grupos en los índices de dificultad de los ítems que funcionan diferencialmente ($d_1 = 0.1$, $d_2 = 0.19$, $d_3 = 0.34$, $d_4 = 0.43$).

b) Nivel de significación utilizado en la determinación de modelo loglineal que mejor se ajusta a los datos ($\alpha = 0.05$ y $\alpha = 0.001$).

c) Tamaño de la muestra (75, 200, 500 y 1000 sujetos por grupo).

Tenemos por tanto cuatro condiciones, una para cada tamaño de muestra. Dentro de cada condición se llevaron a cabo 100 replicaciones para obtener resultados estables.

La longitud del test se fijó en 40 ítems, de los cuales el 10 % presentaban sesgo. Los puntos de corte que determinan el índice de dificultad de los distintos ítems fueron elegidos, siguiendo a Lim y Drasgow (1990), dentro de un rango razonable entre −2 y + 2, de tal forma que la distribución de las puntuaciones de los sujetos en el test se aproximen a una distribución normal. Los índices de dificultad (ID) de la totalidad de los ítems que componen el test se presentan en la Tabla 1.

TABLA 1. VALORES PARAMÉTRICOS DE LOS ID DE LOS ÍTEMS, JUNTO CON LAS CORRESPONDIENTES PUNTUACIONES Z. LOS ÍTEMS SEÑALADOS PRESENTAN DIF, SIENDO SUS ID EN EL GRUPO 2: ÍTEM 1 = 0.40 (Z = 0.25); ÍTEM 2 = 0.31 (Z = 0.50); ÍTEM 3 = 0.16 (Z = 1); ÍTEM 4 = 0.07 (Z = 1.5).

Ítem	*z*	*ID*	*Ítem*	*z*	*ID*	*Ítem*	*z*	*ID*	*Ítem*	*z*	*ID*
1*	0.0	0.50	11	−1.0	0.84	21	0.0	0.50	31	−1.0	0.84
2*	0.0	0.50	12	1.0	0.16	22	0.0	0.50	32	1.0	0.16
3*	0.0	0.50	13	1.5	0.07	23	0.0	0.50	33	1.5	0.07
4*	0.0	0.50	14	−1.5	0.93	24	0.0	0.50	34	−1.5	0.93
5	0.5	0.31	15	1.5	0.07	25	0.5	0.31	35	1.5	0.07
6	−0.5	0.69	16	−1.5	0.93	26	−0.5	0.69	36	−1.5	0.93
7	0.5	0.31	17	−2.0	0.97	27	0.5	0.31	37	−2.0	0.97
8	−0.5	0.69	18	2.0	0.03	28	−0.5	0.69	38	2.0	0.03
9	−1.0	0.84	19	−2.0	0.97	29	−1.0	0.84	39	−2.0	0.97
10	1.0	0.16	20	2.0	0.03	30	1.0	0.16	40	2.0	0.03

Generación de datos

Se generaron vectores de respuestas ($I_i, \ldots, I_n$), para diferentes tamaños de muestra, en un test compuesto por cuarenta ítems dicotómicos a partir de un modelo unifactorial estricto. El modelo es el que sigue:

$$X_i = a_{i1} * F_1 + u_i * E_i \qquad (i = 1, \ldots n); \; [a_{i1}^2 + u_i^2 = 1]$$

siendo el término «a» la saturación factorial de ítem «i» en el factor común F_1 , y el término «u» la saturación factorial del ítem «i» en el factor único E_i, considerado como término error en el modelo. Las puntuaciones factoriales (F_1) y las puntuaciones error ($E_i, \ldots, E_n$) son muestras independientes obtenidas de una distribución normal N(0 1). Dado que tanto las puntuaciones factoriales como las error son independientes, las puntuaciones obtenidas por los sujetos en cada ítem ($X_i, \ldots, X_n$) también se distribuirán normalmente N(0 1).

Las respuestas a los ítems $I_1, \ldots, I_n$ se obtienen dicotomizando las puntuaciones obtenidas en las variables $X_1, \ldots, X_n$ de la siguiente forma:

$$I_i = 0 \text{ si } X_i \leq z \qquad \text{e} \qquad I_i = 1 \text{ si } X_i > z$$

Siendo los puntos de corte elegidos (z, puntuaciones típicas) una función del índice de dificultad de los ítems. Así

$$1 - \text{ID} = P(Z \leq z)$$

De esta forma podemos manipular los índices de dificultad de los diferentes ítems que componen el test.

Siguiendo el procedimiento especificado simulamos los datos para dos muestras independientes de sujetos, siendo la misma estructura factorial para todos los ítems en ambos grupos, y manipulando el índice de dificultad que diferirá, para los ítems con DIF, entre los grupos. La estructura factorial utilizada es

$$X_i = 0.80 * F_1 + 0.60 * E_i$$

La generación de los datos de acuerdo con el modelo y mediante el procedimiento expuesto se hizo dentro del paquete estadístico SPSS/PC + v4.0, a partir de la posibilidad que brinda de generar distribuciones normales.

Análisis

Los modelos lineales logarítmicos permiten determinar las relaciones existentes entre una serie de variables categóricas representadas en tablas de contingencia multidimensionales. Este tipo de análisis especifica los parámetros que representan las propiedades de las variables categóricas y sus relaciones mediante la descomposición lineal de los logaritmos naturales de las frecuencias esperadas en una tabla de contingencia (Gilbert, 1981; Knoke y Burke, 1980; Pardo y San Martín, 1994; Reynols, 1977). En el modelo saturado los componentes que definen la parcelación incluyen todos los efectos principales y las posibles interacciones entre las variables consideradas. En nuestro caso tenemos una tabla de contingencia multidimensional del tipo R × G × H, con frecuencias f_{jkm}, siendo las variables consideradas

R: 2,..., j. Diferentes alternativas del ítem (j = 2).
G: 2,..., k. Grupos comparados (k = 2).
H: 2,..., m. Nivel de habilidad que manifiesta el sujeto que contesta al ítem (m = 8).

La puntuación total del sujeto en el test es tomada como un índice del nivel de competencia del sujeto, y en función de ella los sujetos son asignados a cada uno de los diferentes niveles de habilidad establecidos (ocho

intervalos de amplitud cinco). Los datos obtenidos en la simulación son reordenados en función de las variables señaladas y sometidos a análisis logarítmicos lineales para ver cuál es el modelo que mejor se ajusta a los datos. Para ello se utilizó la opción *Backward* dentro del comando «hiloglinear» del programa estadístico SPSS/PC + v4.0. El procedimiento *Backward*, partiendo del modelo saturado, va eliminando parámetros y comprobando si el efecto de estos parámetros es estadísticamente significativo, es decir, si su presencia contribuye a mejorar el ajuste del modelo. Si un parámetro no es estadísticamente significativo es eliminado del modelo. Así se procede sucesivamente hasta conseguir un modelo en el que todos los términos de orden superior sean estadísticamente significativos ($\alpha = 0.05$ y $\alpha = 0.001$, en este estudio). Los modelos testados de esta forma son todos modelos jerárquicos. Se denominan modelos jerárquicos aquellos en los que la inclusión de los términos de más alto orden implica necesariamente la inclusión de los términos de orden menor que forman parte de ellos. Por ejemplo, si en el modelo está incluida la interacción R × H, necesariamente formarán parte del modelo los parámetros referidos a los efectos de las variables R y H por separado. Para describir un modelo jerárquico es suficiente con enumerar los términos de orden superior que lo componen, a esto se denomina clase generadora del modelo. En nuestro caso, concluiremos que existe DIF si en la clase generadora del modelo que mejor ajusta está presente la interacción entre la respuesta al ítem y el grupo. Es decir, si para el mismo nivel de habilidad existen diferencias entre los grupos en el número de sujetos que responden correctamente el ítem.

Resultados

Se ha tomado como medida de la eficacia de la técnica utilizada tanto el porcentaje de ítems con DIF correctamente detectados, como el porcentaje de ítems sin DIF incorrectamente detectados (falsos positivos). Para el cálculo

TABLA 2. PORCENTAJES DE ÍTEMS CON DIF CORRECTAMENTE IDENTIFICADOS Y DE FALSOS POSITIVOS.

	$\alpha = 0.05$				$\alpha = 0.001$			
Tamaño de muestra	75	200	500	1000	75	200	500	1000
Falsos positivos	6.5	11	25	45	0.22	0.22	4.11	13.67
Ítems con DIF								
4	100	100	100	100	100	100	100	100
3	100	100	100	100	92	100	100	100
2	72	98	100	100	19	77	100	100
1	12	44	67	95	0	3	29	80
Total	71	85.5	91.75	98.75	52.75	70	82.25	95

de estos porcentajes (P = (n.° detecciones/n.° replicaciones)100) se han utilizado los cuatro ítems con DIF junto con nueve ítems que no presentan DIF, uno por cada uno de los diferentes ID utilizados en el estudio. Por ejemplo, los porcentajes de detecciones correctas para cada ítem con DIF será P = (n.° detecciones/100)100, y el cálculo de falsos positivos viene dado por P = (n.° detecciones en los nueve ítems/900)100. Los resultados se presentan en la Tabla 2.

Discusión

El incremento en las tasas de detección conforme aumenta el tamaño de muestra era de esperar. Como es bien sabido cualquier estadístico aumenta su potencia de prueba, permaneciendo igual otras condiciones, conforme aumenta el tamaño de muestra. También era de esperar que los falsos positivos disminuyesen en el nivel de significación más reducido. No era de esperar, sin embargo, un número tan elevado de ítems incorrectamente detectados (25 % y 45 % para 500 y 1000 sujetos; $\alpha = 0.05$). Por otro lado, las tasas de detección de los ítems con más DIF ($d_4 = 0.43$ y $d_3 = 0.34$) son muy altas con independencia del tamaño de muestra y el nivel de significación (mínimo 92 % para $n = 75$ y $\alpha = 0.001$). El ítem 2 que presenta DIF moderado ($d_2 = 0.19$) también presenta un alto porcentaje de detecciones correctas, excepto para $n = 75$ y $\alpha = 0.001$; mientras que con el ítem que presenta menos DIF se necesita un tamaño de muestra de 1000 sujetos por grupo para encontrar elevadas tasas de detección (95 % y 80 %). Sin embargo, la bondad de un procedimiento en la detección del DIF es función de la eficacia que manifiestan en las mismas condiciones procedimientos alternativos. No obstante, comparar los resultados obtenidos por diferentes métodos bajo condiciones similares, sobre todo si éstos son coincidentes, nos puede proporcionar indicios sobre la eficacia relativa de los mismos. Así, Swaminathan y Rogers (1990) en un estudio de simulación para comprobar la eficacia del MH frente a la regresión logística (RL), encontraron unas tasas de falsos positivos en estos procedimientos del 1 % (MH) y del 4 % (RL), y un porcentaje de detecciones correctas del 96 % (MH) y del 94 % (RL). Estas dos pruebas se aplicaron, con un nivel de significación de 0.05, a los datos obtenidos en 20 replicaciones para un test de 80 ítems de los cuales el 10 % presentaba DIF uniforme moderado o alto, con 500 personas por grupo. Los modelos loglineales, en las condiciones más similares a las del estudio referido, ofrecen unas tasas de falsos positivos del 25 % ($\alpha = 0.05$) y del 4.11 % ($\alpha = 0.001$), en tanto el procentaje de ítems correctamente detectados es del 100 %. La comparación de nuestros resultados con los de otro estudio (López Pina, Hidalgo, Sánchez Meca, 1993) en el que se estableció la tasa de error tipo I y la potencia de prueba de los estadísticos chi-cuadrado de Scheuneman, chi-cuadrado de Pearson y el MH, arrojan datos similares. El porcentaje de falsos positivos que se produce al utilizar las pruebas mencionadas era mucho más bajo que el encontrado al utilizar los modelos

loglineales. La tasa de detecciones correctas es sin embargo muy parecida. Parece claro, con los reparos ya expresados por la no equivalencia entre las condiciones en las que se prueban los diferentes métodos, que los modelos loglineales, aunque presentan una aceptable potencia de prueba (detecciones correctas), también tienen unas tasas de error tipo I (falsos positivos) muy altas. Ello en parte puede deberse a que la puntuación total en el test utilizada para agrupar a los sujetos en categorías es calculada incluyendo tanto los ítems con DIF como sin DIF. La puntuación total en el test utilizada como criterio de bloqueo es una función de todas las habilidades que mida el test, y en la medida que se incluyan ítems con DIF, es decir, ítems que discriminan en más habilidades de las que pretende medir el test, el test será multidimensional (Ackerman, 1992; Camilli, 1992; Oshima y Miller, 1992; Mellenbergh, 1989; Shealy y Stout, 1993). El nivel de habilidad estimado a los sujetos está de esta manera distorsionado. Precisamente para conseguir que en lo posible el nivel de habilidades estimado a los sujetos sea unidimensional, se han establecido procedimientos de estimación en etapas. Así, Lord (1980), dentro del marco de la TRI, propuso un esquema para evaluar el DIF en etapas. Consistiría, *grosso modo*, en evaluar el DIF, en un último paso, utilizando las estimaciones de los parámetros del modelo calculadas sobre los ítems que en un primer análisis no presentaron DIF. Otros autores han seguido esta sugerencia utilizando técnicas de evaluación no derivadas de la TRI como el método MH o modelos logit, presentando siempre resultados mejores que cuando el nivel de habilidad de los sujetos no se calculó sobre el test purgado (Mazor, Clauser y Hambleton, 1992; van der Flier, Mellenbergh, Adèr y Wijn 1984; Kok, Mellenbergh y van der Flier, 1985).

Otra variable a considerar es la presencia de grupos con diferentes distribuciones en la habilidad medida. Es decir, ítems que presentan impacto. Estudios como el de Mazor, Clauser y Hambleton (1992) muestran un peor comportamiento del estadístico MH cuando los grupos presentan distribuciones diferentes (tasas de detección 30 %, con igual distribución y 25 % con distintas distribuciones; n = 2000). Habría, por tanto, que determinar la eficacia de los modelos loglineales con distribuciones diferentes y DIF no uniforme, además de comparar su eficacia relativa frente a otras técnicas. Como ventajas *a priori* puede destacarse la posiblidad que brindan, frente a técnicas como el MH o chi-cuadrado, de establecer las relaciones existentes entre un número elevado de variables, siendo quizá en el estudio del comportamiento diferencial de los distractores donde sus posibilidades sean máximas (Green, Crone y Folk, 1989). Otra ventaja adicional de los modelos loglineales es la posibilidad que ofrecen de ser utilizados en la detección del DIF no uniforme. El DIF no uniforme se inferiría si el parámetro del modelo que representa la interacción entre la respuesta al ítem, el grupo de pertenencia y el nivel de habilidad (R × G × H), es estadísticamente significativo, es decir, si contribuye de manera significativa al ajuste del modelo a los datos.

Señalar, para concluir, que a la vista de los resultados obtenidos, no parece muy aconsejable la utilización de los modelos loglineales en la detección del DIF uniforme frente al MH o los métodos chi-cuadrado, ya que estos últimos

son al menos tan eficaces como los modelos loglineales y además no son costosos computacionalmente. Frente a ellos los modelos loglineales presentan la ventaja de constituir un procedimiento válido en la detección del sesgo no uniforme, aunque recientemente Mazor, Clauser y Hambleton (1994) han propuesto una modificación del procedimiento MH que permite detectar este tipo de DIF. Se hacen necesarios, por tanto, estudios que determinen las posibilidades y limitaciones que estos dos procedimientos tienen en la detección del DIF no uniforme.

REFERENCIAS

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias, en R.A. Berk (Ed.), *Handbook of methods for detecting test bias.* Baltimore, M.D.: The Johns Hopkins University Press.

Angoff, W.H. & Ford, S.F. (1973). Item-race interaccion on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95-105.

Angoff, W.H. & Sharon, A.L. (1974). The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement, 34*, 807-816.

Camilli, G. (1979). A critique of the chi-square method of assessing item bias. *Laboratory of Educational Research,* University of Colorado, Boulder.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16* (2), 129-147.

Gilbert, G.N. (1981). *Modelling Society. An introduction to loglinear analysis for social researchers.* London: George Allen & Unwin.

Green, B.F., Crone, C.R. & Folk, V.G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement, 26* (2), 147-160.

Hambleton, R.K. & Rogers, H.J. (1989). Detecting potentially biased test items: Comparison of IRT area an Mantel-Haenszel methods. *Applied Measurement in Education, 2* (4), 313-334.

Hills, J.R. (1989). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice, 8* (4), 5-11.

Holland, W.P. & Thayer, D.T. (1986). *Differential item functioning and the Mantel-Haenszel procedure* (Tecnical Report No.86-89). Princenton, N.J: Educational Testing Service.

Holland, W.P. & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. En H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145).

Hoover, H.D. & Kolen, M.J. (1984). The reliability of six item bias indices. *Applied Psychological Measurement, 8*, 173-181.

Jensen, A.R. (1980). *Bias in mental testing.* New York: Free Press.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54* (4), 681-697.

Kelderman, H. & Macready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27*, 307-327.

Kim, S. & Cohen, A.S. (1991). A comparison of two areas measures for detecting differential item functioning. *Applied Psychological Measurement, 15* (3), 269-278.

Knoke, D. & Burke, P.J. (1980). *Log-linear models.* Beverly Hills, California: SAGE.

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

Lim, R.G. & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75* (2), 164-174.

Linn, R.L, Levine, M.V., Hastings, C.N. & Wardrop, J.L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*, 59-173.

López Pina, J.A., Hidalgo, M.D. y Sánchez Meca, J. (1993). Error tipo I de las pruebas chi-cuadrado en el estudio del sesgo de los ítems. Comunicación presentada al III Simposium de Metodología de las Ciencias del Comportamiento. Santiago de Compostela.

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

Marascuilo, L.A. (1981). Statistical procedures for identifying possible sources of item bias based on chi-square statistics. *Journal of Educational Measurement, 18*, 229-248.

Marascuilo, L.A & Slaugther (1981). Statistical procedures for identifying possible sources of item bias based on χ^2 statistics. *Journal of Educational Measurement, 18* , 229-248.

Mazor, K.M., Brian, E.C. & Hambleton, R.K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*, 443-451.

Mazor,K.M., Clauser, B.E. & Hambleton, R.K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54* (2), 284-291.

Mellenbergh, G.J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105-107.

Mellenbergh, G.J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127-143.

Muñiz, J. (1990). *Teoría de respuesta a los ítems.* Madrid: Pirámide.

Oshima, T.C. & Miller, M.D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16* (3), 237-248.

Osterlind, S.J. (1983). *Test item bias,* Beverly Hills, CA: SAGE.

Pardo, A. y San Martín, R. (1994). *Análisis de datos en psicología II.* Madrid: Pirámide.

Raju, N.S. (1988). The area between two item characteristics curves. *Psychometrika, 53* (4), 495-502.

Raju, N.S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14* (2), 197-207.

Reynols, H.T. (1977). *The analysis of cross-classifications.* NY: Free press.

Rogers, H.J. & Hambleton, R.K. (1989). Evaluation of computer simulated baseline statistics for use in item bias studies. *Educational and Psychological Measurement, 49*, 355-369.

Shealy, R. & Stout, W. (1993). An item response theory model of test bias and differential test functioning. In W.P. Holland & H. Wainer (Eds.). *Differential item functioning* (pp. 197-240). Hillsdale, NJ: LEA.

Spray, J. & Carlson, J. (1986). Comparison of loglinear and logistic regression models for detecting changes in proportions. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Swaminathan, H. & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures, *Journal of Educational Measurement, 27* (4), 361-370.

Westers, P. & Kelderman, H. (1991). Examining differential item functioning due to item dificulty and alternative attractiveness. *Psychometrika, 57* (1), 107-118.

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