Avances en Medición, 3, 29-52
2005

DETECCIÓN DE SESGO EN LOS ÍTEMS MEDIANTE ANÁLISIS DE TABLAS DE CONTINGENCIA

Aura-Nidia Herrera[*]
Universidad Nacional de Colombia, Colombia

Juana Gómez-Benito[**]
Universidad de Barcelona, España

María Dolores Hidalgo-Montesinos[***]
Universidad de Murcia, España

Resumen

El sesgo en los instrumentos de medición psicológica es un tema que ha suscitado tanta polémica como trabajos de investigación en las últimas cuatro décadas, como resultado se dispone hoy de una gama de propuestas de procedimientos para detectar aquellos ítems que tienen funcionamiento diferencial entre grupos de género, raza, etnia, idioma o cualquier otra variable irrelevante para los propósitos de la prueba. Una categoría de estos procedimientos se basa en el análisis de las tablas de contingencia que se pueden conformar a partir de tres variables: una medida de la magnitud del atributo, el grupo al que pertenece el examinado y la puntuación en el ítem de interés. El objetivo de este trabajo es presentar una revisión de los principales procedimientos que se han propuesto hasta el momento desde esa perspectiva, así como sus ventajas y limitaciones en las aplicaciones prácticas. Con base en ésta, se brindan algunas sugerencias y recomendaciones para los potenciales usuarios de las técnicas revisadas.

Palabras clave: *Sesgo en los ítems, DIF, tablas de contingencia, Mantel-Haenszel, Regresión logística, modelos loglineales*

Abstract

Bias in psychological measurement has provoked as much controversy as research during the last four decades. As a consequence, a variety of procedures for detecting items which show differential item functioning between groups of sex, race, culture, language or other irrelevant variables are readily available nowadays. A class of these procedures is based on the analysis contingency tables which comprise three variables: a measure of the ability level, the group of the examinee and the item

[*] Laboratorio de Psicometría. Departamento de Psicología. Universidad Nacional de Colombia. Ciudad Universitaria. Bogotá – Colombia. e-mail: anherrerar@unal.edu.co

[**] Departamento de Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Universidad de Barcelona. Passeig Vall d'Hebrón, 171. 08035 Barcelona-España. e-mail: juanagomez@ub.es

[***] Departamento de Psicología Básica y Metodología. Facultad de Psicología. Universidad de Murcia. Apdo 4021, 30080, Murcia, España. e-mail: mdhidalg@um.es

response. This paper presents a review of these methods, its advantages and limitations in practical works. Finally, some suggestions and recommendations are presented for potential users of the revised techniques.

Key words: *Item bias, DIF, contingency tables, Mantel-Haenszel, logistic regression, loglineal methods.*

Introducción

Desde que en la jerga psicométrica se adoptaran términos como "sesgo", "funcionamiento diferencial de los ítems" (popularizado como DIF, por su abreviatura en inglés) y "funcionamiento diferencial de los tests" (DTF), la producción académica sobre el tema ha ido en permanente ascenso. En un estudio bibliométrico que cubrió la producción de artículos científicos publicados en el último cuarto del siglo XX, Gómez Benito, Hidalgo Montesinos, Guilera Ferré & Moreno Torrente (2005) encontraron que más del 80% de tales publicaciones se hicieron en la última década (1991 a 2000) y dentro de éstos casi las tres cuartas partes se publicaron en el último quinquenio. Este ascenso no resulta sorprendente si se entiende desde una perspectiva no sólo académica sino también social y política: la preocupación cada vez más generalizada por garantizar la igualdad de oportunidades y el tratamiento equitativo de los individuos y grupos sociales, dentro de las diferencias individuales y culturales. Lo que puede resultar sorprendente es que todavía algunas entidades responsables de procesos de evaluación, cuyos resultados puedan tener implicaciones en la asignación de cupos u oportunidades educativas o laborales, no hayan incorporado dentro de sus procesos una estrategia para detectar el posible sesgo en los instrumentos que diseñan o utilizan. Este trabajo presenta una revisión de una clase de tales estrategias: las que se basan en el análisis de tablas de contingencia.

Los trabajos de Eelles, Havighurst, Herrick & Tyler (1951) y de Jensen (1969) se han ganado el calificativo de pioneros en el tema del sesgo en las pruebas psicológicas[1], el primero por mostrar empíricamente que algunos ítems de pruebas de inteligencia se dejaban afectar por diferencias culturales, y el segundo por la enorme polémica que desató en torno a la explicación (genética vs. ambiental) de las diferencias individuales evidenciadas por las pruebas; sin embargo, varias décadas antes Stern (1914) había mostrado que el rendimiento en pruebas de inteligencia variaba según la clase social y Binet & Simon (1916) habían eliminado algunos ítems en la nueva versión de su prueba de inteligencia porque eran sensibles a efectos culturales. Hoy, tras las acaloradas discusiones de los sesenta -muchas de las cuales se adelantaron fuera de la esfera académica- después del trabajo de Jensen (1980) que buscaba despejar el término "sesgo" de las connotaciones éticas, sociales y políticas para entenderlo

[1] Algunas revisiones históricas y conceptuales sobre el tema pueden encontrarse en Angoff (1993), Camilli & Shepard (1994), Fidalgo (1996a), Gómez Benito & Hidalgo Montesinos (1997b), Ferreres (1998) y Herrera, Sánchez Pedraza & Gómez Benito (2001), entre otros.

como un problema técnico, y de la oportuna propuesta de Holland & Thayer (1988) para introducir la expresión Funcionamiento Diferencial de los Ítems (DIF) como un problema diferente al controvertido "sesgo", parece haberse llegado a un relativo acuerdo sobre los términos y haberse logrado un importante desarrollo tecnológico y metodológico para cuidar que los instrumentos de medición psicológica funcionen de manera similar en grupos de género, raciales, culturales o étnicos diferentes. Debe anotarse, sin embargo, que el relativo acuerdo sobre los términos no supuso superar la preocupación inicial sobre las diferencias entre estos grupos, en cuanto al rendimiento en las pruebas y, por ende, en las oportunidades educativas y laborales; prueba de esto puede encontrarse en los más recientes escritos de Jensen (1998); Jensen (2000) o el número especial de la revista *Inteligencia* editado por Gottfredson (1997). De hecho, el sesgo y el DIF siguen considerándose temas "candentes" dentro de la psicometría (Gómez-Benito & Hidalgo-Montesinos, 2003).

En palabras de Muñiz (1998) "un metro estará sistemáticamente sesgado si no proporciona la misma medida para dos objetos o clases de objetos que de hecho miden lo mismo, sino que sistemáticamente perjudica a uno de ellos" (p.236). Pero el metro puede hacer referencia bien a una prueba como un todo, o bien a los elementos que la componen; el énfasis de la producción académica de los últimos años ha estado en los segundos y este trabajo versa también sobre el DIF. Se entiende que un ítem funciona diferencialmente, o tiene DIF, cuando su puntuación varía en función de alguna variable como raza, género, etnia, idioma, grupo cultural, etc., que es irrelevante para el constructo psicológico que pretende medir la prueba. Visto así, el reto para la psicometría ha sido el diseño de estrategias que permitan comparar la probabilidad de acierto de un ítem entre grupos iguales en la magnitud del atributo medido; esta aproximación tiene varias implicaciones metodológicas. La primera y más evidente es que la detección de DIF supone una comparación de grupos, generalmente dos, uno denominado focal que se considera desfavorecido y que frecuentemente es minoritario, y otro conocido como de referencia por cuanto se considera un estándar de comparación del grupo focal. Las curvas ascendentes de la figura 1 representan la probabilidad de acierto (eje *y*) en tres ítems diferentes en función de la magnitud del atributo expresada en una escala de media *0* y desviación estándar *1* (eje *x*) para dos grupos. Puede observarse que mientras en el ítem *i* igual magnitud del atributo corresponde a igual probabilidad de acierto para los dos grupos, en los otros dos ítems no ocurre lo mismo; estos últimos presentan DIF. Sin embargo, la probabilidad de acierto en el ítem *j* es inferior para el grupo 2 en todos los valores de *x* mientras que para el ítem *k* la probabilidad es menor para el grupo 1 cuando *x* es inferior a 1, pero esto cambia cuando *x* es alta. El ítem *j* tiene DIF uniforme en contra del grupo 2 mientras que el ítem *k* tiene DIF no uniforme en contra del grupo 1 cuando la magnitud del atributo es baja, y en contra del grupo 2 cuando la magnitud del atributo es alta.

De otra parte, el solo hecho de que un instrumento de medida arroje resultados sistemáticamente inferiores para un grupo en comparación con otro no constituye evidencia de sesgo, ya que si efectivamente existen diferencias entre los grupos en lo que la prueba mide es apenas de esperarse que sus resultados las muestren. Estas diferencias se conocen en el lenguaje técnico como impacto o diferencias válidas. En la figura 1 también se representan las distribuciones de la magnitud del atributo, mientras que para el ítem *j* esta distribución es igual para los dos grupos, el grupo 1 tiene en promedio, mayor magnitud del atributo medido por el

ítem *i* que el grupo 2 y éste último tiene, en promedio, mayor magnitud del atributo medido por el ítem *k*. En resumen, el ítem *i* presenta impacto y no tiene DIF, el ítem *j* tiene DIF uniforme y no presenta impacto y, finalmente el ítem *k* tiene DIF no uniforme e impacto.

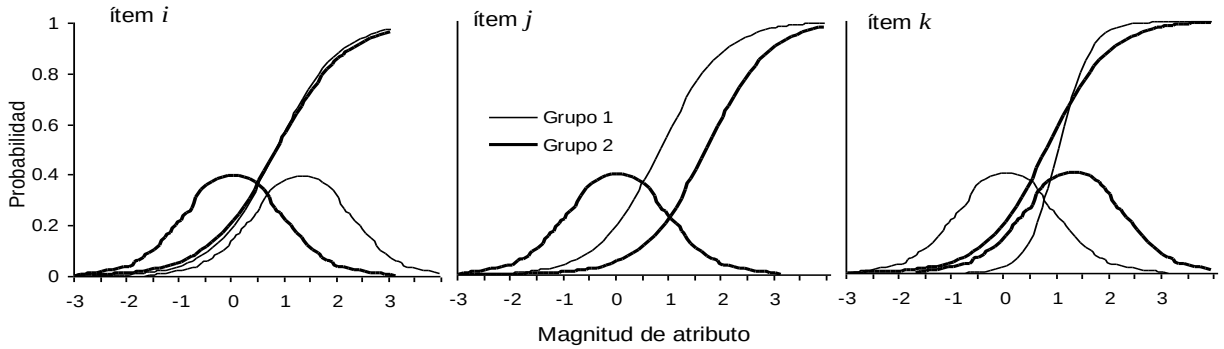

Figura 1. Probabilidad de acierto en función de la magnitud del atributo (curvas ascendentes) y distribución de la magnitud del atributo medido por tres ítems en dos grupos.

Para la psicometría ha constituido un verdadero reto diseñar procedimientos que no confundan DIF e impacto comparando individuos o subgrupos iguales en cuanto a la magnitud del atributo objeto de la prueba. Si se dispone de un criterio válido externo a la prueba (otro metro no sesgado) resulta sencillo equiparar los grupos y comparar su probabilidad de acierto en el ítem, pero en ausencia de dicho criterio, que es la situación usual en la práctica, la única evidencia empírica disponible es el vector de respuestas del examinado al instrumento, el cual incluye los ítems presuntamente sesgados. El problema tiene entonces una circularidad que debe romperse y que constituye lo que Fidalgo, (1996a) denomina la paradoja de los procedimientos para evaluar DIF, puesto que conduciría a que "sólo es posible evaluar correctamente el DIF cuando es innecesario" (p. 435).

Buena parte de la producción de las últimas décadas ha estado dedicada a proponer procedimientos que satisfagan esas exigencias metodológicas y que sean viables y efectivos en la práctica. Camilli & Shepard (1994) clasifican tales propuestas en tres categorías: a) los métodos basados en el análisis de varianza y en la Teoría Clásica de los Test (TCT), b) los que se basan en la Teoría de Respuesta al Ítem (TRI) y c) los que se basan en el análisis de tablas de contingencia (TC). En Ferreres (1998) se puede encontrar una revisión de algunas otras clasificaciones que obedecen a criterios diferentes. Siguiendo la clasificación de Camilli & Shepard (1994), este trabajo se ocupa de los métodos basados en el análisis de TC. Los métodos incluidos en su primera categoría ya están en desuso puesto que fallaron en la equiparación de los grupos y en consecuencia no lograron distinguir DIF de impacto; por su parte, los métodos basados en la TRI gozan de una exquisita fundamentación matemática y tienen muchas fortalezas metodológicas pero sus exigencias en cumplimiento de supuestos y tamaños de muestra los hacen poco aplicables en algunas condiciones prácticas; por el contrario, la tercera categoría de métodos pueden resultar más intuitivos, sencillos, de bajo costo computacional y menos exigentes, por lo que se constituyen en una opción viable para los constructores de pruebas en condiciones reales poco favorables.

Dentro de los métodos basados en el análisis de tablas de contingencia se pueden distinguir dos enfoques: los que se fundamentan en la prueba de hipótesis sobre la igualdad de proporciones analizando tablas de contingencia bidimensionales, y los que generan modelos para el análisis de variables categóricas con tablas de más de dos dimensiones. Dentro de los primeros se encuentran algunas aplicaciones de la prueba χ^2, el método de estandarización y el Mantel-Haenszel (MH); mientras que dentro de los últimos se pueden incluir los modelos logit y log-lineales y la regresión logística (RL). Aquí se revisan estas técnicas haciendo especial énfasis en el MH y en la RL, dos de los más frecuentemente utilizados.

Comparación de proporciones

Los métodos que comparan proporciones se fundamentan en que el DIF se detecta probando una hipótesis sobre la igualdad de proporciones entre los grupos igualados por la magnitud del atributo medido. Esta última, tomando como medida el puntaje total en la prueba, se fracciona de manera que genere diferentes estratos ($K = 1,..k,...m$) y para cada estrato se construye una tabla de contingencias que cruza las variables grupo ($J = 1,..j,...g$) y respuesta en el ítem ($I = 1,...i...p$). Así, cuando se trata de ítems dicótomos ($I = 0,1$) y se comparan solamente dos grupos ($J = r, f$), la información completa sobre los aciertos y fallos en cada ítem tiene tantas tablas bidimensionales de *2 x 2,* como estratos de la magnitud del atributo medido; cada una de ellas tiene la estructura que se muestra en la tabla 1,

Tabla 1
Estructura de una tabla de contingencia correspondiente a un nivel *k* de magnitud del atributo

Puntaje en el ítem	Grupo		Total
	Referencia	Focal	
1	f_{1rk}	f_{1fk}	f_{1k}
0	f_{0rk}	f_{0fk}	f_{0k}
Total	f_{rk}	f_{fk}	f_k

donde: f_{1rk} es el número de examinados de magnitud del atributo *k* y del grupo de referencia que acertaron en el ítem

f_{1fk} es el número de examinados de magnitud del atributo *k* y del grupo focal que acertaron en el ítem

f_{0rk} es el número de examinados de magnitud del atributo *k* y del grupo de referencia que fallaron en el ítem

f_{0fk} es el número de examinados de magnitud del atributo *k* y del grupo focal que fallaron en el ítem

f_{rk} y f_{fk} son el número total de examinados de los grupos de referencia y focal, respectivamente, con magnitud del atributo *k*.

f_{1k} y f_{0k} son el número total de examinados con magnitud del atributo *k* que acertaron y fallaron en el ítem, respectivamente

f_k es el número total de examinados de magnitud del atributo *k*.

Aplicaciones de χ^2

Un procedimiento propuesto por Scheuneman (1979), y conocido posteriormente como el *Ji* cuadrado de los aciertos, parte del supuesto de que se puede descartar la presencia de DIF en el ítem si dentro de cada estrato la proporción de aciertos es igual en cada uno de los grupos. En este sentido, propuso probar la hipótesis de igualdad de proporciones de los aciertos mediante un estadístico χ^2 con $(m-1)(g-1)$ grados de libertad, siendo m el número de estratos y g el número de grupos. El estadístico, como la prueba Ji cuadrado tradicional, se basaba en la comparación de las frecuencias observadas y las esperadas en cada una de las celdas de tablas de contingencias, como la de la tabla 1, pero considerando solamente las proporciones de acierto en cada ítem. Así, si f_{1jk} y e_{1jk} son la frecuencia observada y esperada, respectivamente, de personas del grupo *j* y del estrato *k* que aciertan al ítem, el estadístico se define como $\chi^2_{aciertos} = \sum_{j=1}^{g}\sum_{k=1}^{m}\frac{(e_{1jk} - f_{1jk})^2}{e_{1jk}}$

Aunque la sencillez y economía del procedimiento habrían podido augurar un uso generalizado, Baker (1981) y el mismo Scheuneman (1981) discutieron cómo este procedimiento, al considerar solamente los aciertos, puede inestabilizarse cuando existe impacto, caso en el cual los resultados pueden depender de qué tan similares sean los tamaños de los dos grupos; en estas condiciones, no resulta claro que la distribución que siga sea realmente la χ^2. De acuerdo con Ironson (1982), Camilli propuso una modificación al procedimiento considerando tanto las proporciones de acierto como las de los fallos; esta suma seguiría una distribución χ^2 con $m(g-1)$ grados de libertad. De esta manera, siguiendo la misma notación utilizada hasta ahora, **el χ^2 completo** se define como $\chi^2_{completo} = \sum_{i=0}^{1}\sum_{j=1}^{g}\sum_{k=1}^{m}\frac{(e_{ijk} - f_{ijk})^2}{e_{ijk}}$

Tanto el procedimiento inicial como su modificación fueron criticados por Mellenbergh (1982) por su incapacidad para detectar DIF no uniforme; además, ambos procedimientos se inestabilizan con valores bajos dentro de las celdas, con diferencias en los tamaños de muestra o ante la presencia de impacto en los grupos. En la actualidad estos procedimientos no se utilizan y no son recomendados por los autores sobre el tema.

Método de estandarización

La primera referencia que generalmente se reporta sobre el método de estandarización, también conocido como Diferencia de Proporciones Estandarizadas (DPE), son las publicaciones de Dorans & Kulick (1986) y Dorans (1989); sin embargo, según Dorans & Holland (1993) los primeros desarrollos del procedimiento se iniciaron tres años antes mediante trabajos de aplicación en diferentes pruebas del Educational Testing Service (ETS)

cuyos resultados se encuentran en Kulick & Dorans (1983a, 1983b) y Dorans & Kulick (1983), reportes de investigación no publicados.

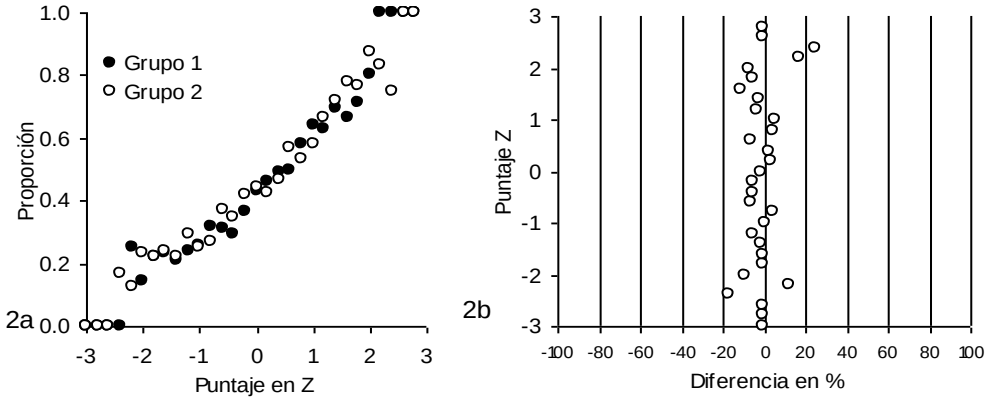

Figura 2: Representaciones gráficas del comportamiento de un ítem sin DIF: 2a: Diagrama de dispersión de la proporción de acierto en función del puntaje en la prueba; 2b: magnitud de las diferencias entre los grupos en función del puntaje en la prueba

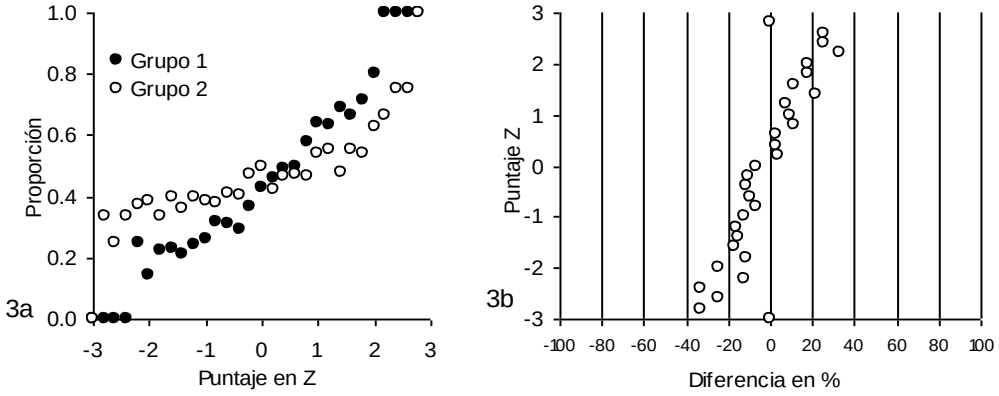

Figura 3: Representaciones gráficas del comportamiento de un ítem con DIF: 3a: Diagrama de dispersión de la proporción de acierto en función del puntaje en la prueba; 3b: magnitud de las diferencias entre los grupos en función del puntaje en la prueba

El procedimiento se basa en la comparación del desempeño del ítem en los grupos de interés cuando se controla el nivel de magnitud del atributo. Inicialmente ese desempeño se estimó a partir de las proporciones condicionales de aciertos en el ítem para los grupos, lo que le dio al procedimiento la denominación DPE; sin embargo, Dorans y Holland (1993) lo estiman por medio de regresiones no paramétricas ítem test para los dos grupos separadamente. La esencia del procedimiento puede entenderse muy intuitivamente a través de una inspección visual como la que aparece en las figuras 2 y 3; éstas muestran el comportamiento de un ítem sin DIF y uno con DIF, respectivamente, aplicados a 592 personas del grupo 1 y 1140 del grupo 2, con idéntica distribución en los puntajes totales en la prueba. En las figuras 2a y 3a aparecen los diagramas de dispersión de las proporciones de acierto en

el ítem en función del rango (de longitud 0,2) de puntaje en la prueba total, expresado en puntuación Z; en las figuras 2b y 3b se muestran las diferencias de proporción entre los dos grupos expresadas en porcentaje, en función de los mismos rangos de puntaje total en la prueba. La comparación de las dos figuras permiten afirmar que en el primer caso (figura 2) el comportamiento del ítem es bastante similar en los dos grupos y las diferencias, que se encuentran en un pequeño rango alrededor del 0, pueden considerarse despreciables; es un ítem que no tiene DIF. Por el contrario, tanto el diagrama de dispersión como las diferencias entre grupos en el ítem de la figura 3 muestran discrepancias importantes y patrones que permiten identificar el tipo de DIF: las curvas para los dos grupos se cruzan alrededor $z = 0.5$ (ver figura 3a) y además, las diferencias son negativas para bajos puntajes en la prueba y positivas para puntajes altos (ver figura 3b); se trata de un ítem que muestra DIF no uniforme.

Además de esta herramienta visual, utilizando las proporciones condicionales, un índice numérico que permite identificar los ítems con DIF es $DPE = \dfrac{\sum_{k=1}^{m} w_k \left(p_{fk} - p_{rk} \right)}{\sum_{k=1}^{m} w_k}$

donde $p_{fk} = \dfrac{f_{1fk}}{f_{fk}}$ y $p_{rk} = \dfrac{f_{1rk}}{f_{rk}}$ son las proporciones de acierto en el ítem para los examinados del estrato *k* en los grupos focal y de referencia, respectivamente, y w_k es un factor de ponderación de la diferencia en dicho estrato.

Este factor de ponderación es común para los dos grupos, lo cual constituye uno de los elementos esenciales del procedimiento ya que diferencia el cálculo del DIF del de impacto, definido como la diferencia de proporciones ponderadas cada una por la frecuencia relativa del grupo correspondiente. La elección del factor de ponderación depende de los objetivos de la investigación. Algunos valores que puede tomar w_k son: a) f_k, el número total de examinados en el estrato *k*; b) f_{rk}, el número de examinados del grupo de referencia en el estrato *k*; c) f_{fk}, el número de examinados del grupo focal en el estrato *k*; o d) la frecuencia relativa en el estrato *k* en algún grupo de referencia. Uno de los que se usa con mayor frecuencia es el número de examinados del grupo focal en el estrato correspondiente (f_{fk}).

Los valores que puede tomar el DPE van entre *–1* y *+1* y si los valores son positivos el ítem está favoreciendo al grupo focal. Además, Dorans & Holland (1993) sugieren rangos de interpretación: valores entre –0.05 y 0.05 se consideran despreciables, valores absolutos entre 0.05 y 0.1 son dudosos y valores por fuera de estos últimos rangos obligan a una revisión cuidadosa de los ítems. Estos valores, sin embargo, se pude transformar a una métrica delta, denominada DPE D-DIF, que arroja correlaciones mayores con el estadístico Mantel-Haenszel, procedimiento que se tratará a continuación. Aunque estos índices numéricos (el DPE y el DPE D-DIF) se venían utilizando en el ETS como uno de los procedimientos para la descripción del DIF, algunos años más tarde se desarrollaron las formas de calcular sus

respectivos errores estándar (Dorans & Holland, 1993), lo que permite someter a prueba la hipótesis de la existencia de DIF con algún nivel de confianza, siguiendo una distribución normal estándar. Una explicación detallada de este procedimiento se encuentra en Fidalgo (1996a).

Mantel Haenszel

El estadístico Mantel-Haenszel (MH), propuesto originalmente por Mantel & Haenszel (1959) para el análisis de grupos pareados en el ámbito de los estudios epidemiológicos y utilizado posteriormente para detección de DIF por Holland & Thayer (1986, 1988), es hoy uno de los procedimientos más populares en el ámbito de los estudios de sesgo en ítems dicótomos. Su éxito se debe, además de su sencillez y economía computacional, a que maneja de manera eficiente los distintos niveles de habilidad como una variable de control, permitiendo evaluar y describir cómo la relación entre las variables grupo y respuesta en el ítem se modifica por la presencia de otra variable categórica con *m* estratos (magnitud del atributo medido). El MH parte del supuesto de que si un ítem no tiene DIF, la razón entre el número de personas que aciertan y fallan el mismo debe ser igual en los grupos de comparación para los *m* estratos. Si, siguiendo la notación utilizada hasta ahora, f_{1rk} y f_{0rk} son el número de examinados del grupo de referencia y del estrato k que acertaron y fallaron en el ítem, respectivamente; y f_{1fk} y f_{0fk} son el número de examinados del grupo focal y del estrato k que acertaron y fallaron en el ítem, respectivamente; la supuesta igualdad de razones puede expresarse como $\frac{f_{1rk}}{f_{0rk}} = \alpha \frac{f_{1fk}}{f_{0fk}}$, lo que permite afirmar que $\frac{f_{1rk} f_{0fk}}{f_{0rk} f_{1fk}} = \alpha = 1$; razón conocida como *Odds ratio* para el estrato k. El MH prueba la hipótesis de que en m estratos de magnitud del atributo medido, la razón (llamada *Odds ratio común* y notada generalmente como α_{MH}), es igual a *1*. Un estimador de este *Odds ratio* común es $\hat{\alpha}_{MH} = \frac{\sum_{k=1}^{m} f_{1rk} f_{0fk} / f_k}{\sum_{k=1}^{m} f_{0rk} f_{1fk} / f_k}$.

Y el estadístico de prueba para $H_0 : \alpha_{MH} = 1$ contra $H_1 : \alpha_{MH} \neq 1$, conocido como el Ji cuadrado de Mantel-Haenszel sigue una distribución χ^2 con un grado de libertad, y está dado por $\chi^2_{MH} = \frac{\left(\left| \sum_{k=1}^{m} f_{1rk} - \sum_{k=1}^{m} E(f_{1rk}) \right| - 0{,}5 \right)^2}{\sum_{k=1}^{m} Var(f_{1rk})}$,

donde: $E(f_{1rk}) = \frac{f_{rk} f_{1k}}{f_k}$ es el número de personas del grupo de referencia y del estrato *k*, que deben acertar en el ítem si éste no tiene DIF; es decir, la frecuencia esperada en la

celda *1rk* calculada a partir del producto de las respectivas marginales, como en las aplicaciones tradicionales de Ji cuadrado, y

$Var(f_{1rk}) = \frac{f_{rk} f_{1k} f_{fk} f_{0k}}{f_k^2 (f_k - 1)}$ es la varianza de la celda *1rk*, también calculada a partir de las marginales de cada tabla de contingencias, asumiendo una distribución hipergeométrica.

Si se rechaza H_0, el ítem estudiado tiene DIF y el valor del estimador $\hat{\alpha}_{MH}$, que puede variar en el rango $(0,\infty)$, da información acerca de la magnitud y dirección del mismo: $\hat{\alpha}_{MH} > 1$ sugiere sesgo a favor del grupo de referencia, mientras que $\hat{\alpha}_{MH} < 1$ sugiere sesgo en contra del mismo. Estos valores, sin embargo, pueden expresarse en métricas diferentes que pueden facilitar la interpretación o comparación de los resultados. Una de éstas métricas propuesta por Holland & Thayer (1985) conocida como escala delta del ETS, análoga a la empleada en el procedimiento de estandarización, es $MH\ \ D-DIF = -2.35 ln(\hat{\alpha}_{MH})$. Esta transformación expresa los valores de $\hat{\alpha}_{MH}$ en una escala simétrica donde 0 es el valor de hipótesis nula, es decir, indica ausencia de DIF, y los valores negativos o positivos informan la dirección y magnitud del DIF en términos de diferencia de la dificultad del ítem. Valores negativos indican que el ítem resulta más fácil para el grupo de referencia (DIF en contra del grupo focal) y valores positivos tienen la interpretación inversa.

Dorans & Holland (1993) muestran otras dos transformaciones que también permiten interpretaciones en términos de la dificultad del ítem. La primera de ellas consiste en una conversión de las proporciones de acierto a una escala *z* usando la inversa de la función normal acumulada, para posteriormente expresarlos en una nueva escala de media 12 y desviación estándar 4, mediante una transformación lineal. La segunda, conocida como métrica *p*, es

$$MH\ \ P-DIF = p_f - P_r$$

donde: $P_r = \frac{\hat{\alpha}_{MH} p_f}{(1 - p_f) + \hat{\alpha}_{MH} p_f}$ puede verse como la proporción de aciertos predicha para el grupo de referencia con base en el *Odds ratio* de Mantel-Haenszel, y

p_f es la proporción de aciertos en el grupo focal.

De otra parte, Robins, Breslow & Greenland (1986) y Phillips & Holland (1987) propusieron expresiones equivalentes para estimar el error estándar del log de Mantel-Haenszel, las cuales permiten obtener un estimador del error estándar para el *MH D-DIF*. Por su parte, Holland (1989) desarrolló una expresión para estimar el error estándar del *MH P-DIF*. Estos, sin embargo no han tenido uso tan generalizado como el estadístico χ^2_{MH} para

identificar DIF y la estimación de α_{MH} para describirlo. La explicación de dichas expresiones pueden encontrarse en Dorans & Holland (1993); además, en Holland & Thayer (1988), Donoghue, Holland & Thayer (1993), Camilli (1993) y Fidalgo, (1996a), entre otros, se encuentran explicaciones detalladas e ilustraciones del procedimiento completo.

Sin lugar a dudas este es uno de los procedimientos más utilizados en los estudios actuales sobre sesgo en ítems de calificación dicótoma. Entre las razones que justifican tal hecho se encuentran, además de su eficiencia en la detección de DIF, su sencillez, su bajo costo computacional y sus bajos requerimientos de tamaño de muestras o supuestos distribucionales. Sin embargo, se han identificado algunas limitaciones, una de las cuales, común a todos los métodos que utilizan los puntajes observados en la prueba como criterio de pareamiento de los grupos, es la posible contaminación de los ítem DIF en dicho criterio. Como respuesta a esta dificultad, Holland & Thayer (1988) propusieron un procedimiento de dos etapas que permite refinar el criterio de pareamiento. En la primera etapa se aplica el procedimiento tal cual se ha descrito, utilizando como criterio el puntaje en la totalidad de los ítems de la prueba y se identifican los ítems DIF. En la segunda etapa se recalifica la prueba excluyendo los ítems identificados en la etapa anterior, exceptuando el que se está analizando, y se repite el procedimiento para todos los ítems; así, en el criterio de pareamiento se ha excluido la posible contaminación de los ítems DIF, exceptuando el ítem que se está estudiando. De otra parte, en su tesis doctoral no publicada, Fidalgo (1996b) propuso un procedimiento iterativo que también se inicia con la aplicación del procedimiento para identificar los ítems DIF y continua iterativamente purificando el criterio de pareamiento mediante la eliminación de los ítems identificados con DIF; el proceso se detiene cuando en dos iteraciones consecutivas se identifiquen como DIF los mismos ítems. Diferentes estudios como los de Miller & Oshima (1992), Clauser, Mazor & Hambleton (1993), Navas-Ara & Gómez Benito (1994, 2002); Fidalgo & Mellenbergh (1995) y Fidalgo (1996b) han mostrado que un procedimiento de purificación de la medida de la magnitud del atributo mejora eficientemente el poder del estadístico MH y, en consecuencia, su uso se ha generalizado hoy.

Otra de las limitaciones que se ha reportado del MH es su incapacidad para detectar DIF no uniforme; los hallazgos de Narayanan & Swaminathan (1996); Rogers & Swaminathan (1993); Swaminathan & Rogers (1990), son bastante consistente al respecto. Esto ocurre especialmente cuando los ítems tienen dificultad media; sin embargo, su poder mejora cuando se trata de ítems de alta o baja dificultad. Puesto que el DIF no uniforme ocurre cuando las curvas se cruzan como ocurre en el ítem *k* de la figura 1, si este cruce ocurre hacia el centro de la escala de magnitud del atributo el MH no identifica este tipo de interacción; por el contrario, si las curvas se cruzan hacia los extremos bajo o alto de la escala de magnitud del atributo, el MH detecta un efecto de grupo (Rogers & Swaminathan, 1993; Uttaro & Millsap, 1994). Con el propósito de superar esta limitación Mazor, Clauser & Hambleton (1994) propusieron una modificación del MH que consiste en correr separadamente el procedimiento para los grupos con mayor y menor desempeño en la prueba y comparar los resultados. Fidalgo & Mellenbergh (1995) han reportado resultados satisfactorios utilizando esta modificación pero los resultados hallados por Hidalgo Montesinos & López Pina (2004) mostraron que al utilizar esta modificación no hay ganancia en la tasa de detección de DIF uniforme y el aumento en las detecciones correctas de DIF no uniforme no es muy importante.

Dada la rápida popularización del MH las dos últimas décadas han sido testigos de la producción de un importante volumen de trabajos que han analizado el efecto de diversos factores sobre su poder, su distribución y su error tipo I. Algunos de los factores estudiados han sido el tamaño de muestra, el número de estratos o rangos que se emplean para separar los estratos de magnitud del atributo, el tipo de ítem estudiado en términos de los valores de sus parámetros, las diferencias en la distribución de magnitud del atributo entre los grupos focal y de referencia, el porcentaje de ítems DIF dentro de la prueba, y el tipo y magnitud de DIF del ítem estudiado. Hambleton, Clauser, Mazor & Jones (1993) presentan un completo resumen de los principales trabajos realizados entre los 80 y primeros años de los 90, y sus implicaciones prácticas más relevantes. Sin duda, uno de los factores que mayor interés despierta por sus implicaciones en la práctica es el tamaño de muestra necesario para obtener resultados satisfactorios. De acuerdo con la revisión de Hills (1990), el MH es una opción adecuada cuando el tamaño de los grupos está entre 100 y 300; para Zieky (1993) se requieren por lo menos 100 examinados en el grupo más pequeño y 500 en total; mientras que Mazor, Clauser & Hambleton (1992) y Fidalgo, Mellenbergh & Muñiz (1999) cuestionan su uso con muestras muy pequeñas. Para los primeros, un tamaño de 200 es recomendable cuando interesa identificar únicamente los ítems con amplia magnitud de DIF; si se requiere mayor poder el tamaño de muestra debe incrementarse de manera importante, sobre todo cuando hay impacto. Por su parte, Fidalgo, Mellenbergh & Muñiz (1999) encontraron que el χ^2_{MH} es más sensitivo al tamaño de muestra que el $\hat{\alpha}_{MH}$ y recomendaron utilizar ambos estadísticos, dando prioridad al segundo cuando se tengan muestran tan pequeñas como 200 examinados por grupo. En estudios más recientes Fidalgo, Ferreres & Muñiz (2004) encontraron que con grupos de referencia entre 50 y 200 examinados y grupos focales entre 50 y 100, el poder del MH es muy bajo y solamente mejora si se utiliza un nivel de significancia del 20% en vez del tradicional 5%. Finalmente, Herrera & Gómez Benito (en revisión) encontraron que con 500 examinados en el grupo de referencia y 100 en el focal, las tasas de detección del MH fueron bajas; sin embargo, éstas mejoraron considerablemente con tamaños de 500 y 125 para DIF uniforme y mixto, pero para DIF no uniforme se obtuvo una tasa apenas aceptable con grupos iguales de 1500 examinados cada uno, esta tasa parece mejorar cuando los ítems tienen dificultad extrema (muy alta o muy baja).

Otro tipo de estudios han comparado el MH con otros procedimientos para evaluar su eficiencia relativa. En dos estudios diferentes Narayanan & Swaminathan (1994) y Roussos & Stout (1996) compararon el comportamiento del error tipo I del MH y SIBTEST de Shealy & Stout (1993) y no observaron grandes diferencias entre ellos. Desde una perspectiva similar Narayanan & Swaminathan (1996) compararon el error tipo I y el poder del MH, la regresión logística (RL) y el SIBTEST y encontraron tasas de detección del DIF uniforme muy similar para los tres procedimientos pero muy inferiores para el MH cuando se trataba de detectar DIF no uniforme. De otra parte Rogers & Swaminathan (1993) evaluaron las propiedades distribucionales y el poder del MH y la RL y encontraron diferencias entre los dos procedimientos para detectar DIF uniforme y no uniforme, pero poder similar para detectar DIF mixto. En el trabajo de Miller & Oshima (1992) el MH mostró comportamiento más estable a través del tamaño de muestra, comparado con seis índices de DIF basados en la TRI. Finalmente, Gómez Benito & Navas-Ara (2000) compararon el poder y el error tipo I del MH

con el de otros seis procedimientos para detección de DIF, incluyendo tres basados en la TRI, y encontraron que el MH mostró los mejores resultados en términos de tasas de detecciones correctas y falsos positivos.

Modelos para el análisis de tablas

A diferencia de los procedimientos descritos hasta ahora, los que se presentan a continuación se fundamentan en el ajuste de modelos que expliquen las relaciones existentes entre las variables consideradas en las tablas de contingencia descritas antes (magnitud del atributo tomando como medida el puntaje total en la prueba, $K = 1,...k,...m$; grupo, $J = r, f$; y respuesta en el ítem $I = 1,0$), o la estimación de los parámetros de un modelo que explique la probabilidad de acierto en el ítem, en función de las mismas variables. Comparar el ajuste de los diferentes modelos, o rechazar la hipótesis para cada uno de los parámetros, brinda información no solamente sobre la presencia sino sobre el tipo de DIF.

Modelos log-lineales y modelos logit

Los modelos log-lineales, modelos lineales sobre el logaritmo natural de las frecuencias esperadas, permiten analizar la interacción entre todas las variables de una tabla de contingencia simultáneamente, incluyendo las interacciones entre grupos de variables; es decir, los modelos log-lineales analizan una tabla de contingencia multidimensional, en lugar de *m* tablas bidimensionales. Dicho brevemente, se trata de ajustar un modelo que busca predecir la frecuencia esperada en cada celda de la tabla como un producto de los efectos (efectos principales y sus interacciones); posteriormente se efectúa una transformación logarítmica para convertir dicho producto en un sistema lineal. Cuando no hay ningún efecto significativo, el valor esperado en cada celda es igual al total de observaciones dividido por el número de celdas; si los totales marginales son diferentes, se debe a la presencia de algún efecto. Si un modelo se ajusta a los datos, es decir, explica de manera satisfactoria las relaciones existentes entre las variables, el valor esperado en cada celda no se diferencia significativamente de la frecuencia observada en la misma; si son diferentes, debe probarse otro modelo. Además, dependiendo de los parámetros que resulten significativos en el modelo ajustado, este tipo de análisis permite describir las características de las variables y sus interacciones, por lo que Mellenbergh, (1982) los propuso como herramientas útiles para la detección de diferentes tipos de DIF. En Knoke & Burke (2000) o en Powers & Xie (2000) se encuentran explicaciones sencillas y completas de este tipo de modelos estadísticos, aquí se hará una breve aproximación de su aplicación a la detección de ítems DIF.

En general, se trata de ajustar modelos donde la variable dependiente es la frecuencia esperada en la celda y, como es bien sabido, bajo la hipótesis nula de independencia de los efectos, la probabilidad de una celda se expresa en términos multiplicativos. Sin embargo, utilizando una transformación logarítmica el modelo puede expresarse mediante un sistema lineal, que para la primera celda ($1rk$) de la tabla 1 tendría la forma:

$$ln(e_{1rk}) = \lambda + \lambda_{I(1)} + \lambda_{J(r)} + \lambda_{K(k)},$$

donde: e_{1rk} es la frecuencia esperada en la celda *1rk* , es decir el número de personas del grupo de referencia y del estrato *k*, que se espera, acierten en el ítem.
$\lambda_{I(1)}$ es el efecto de la respuesta al ítem (variable *I*) en la celda *1* (acierto),
$\lambda_{J(r)}$ es el efecto del grupo (variable *J*) en la celda *r* (grupo de referencia)
$\lambda_{K(k)}$ es el efecto del nivel de magnitud del atributo (variable *K*) en la celda *k* (*k-ésimo* estrato).

Este sistema lineal expresa un modelo de efectos principales pero también pueden generarse modelos de interacciones de segundo, tercer o más órdenes, dependiendo del número de variables. En la detección de DIF un modelo con interacciones de segundo orden incluye las interacciones entre respuesta al ítem y el grupo, respuesta al ítem con magnitud del atributo, y grupo con magnitud del atributo. Un modelo saturado es el que incluye todos los efectos principales e interacciones posibles; es decir, incluye además de las interacciones mencionadas, la de los tres efectos, y se expresa en general como:

$$ln(e_{ijk}) = \lambda + \lambda_{I(i)} + \lambda_{J(j)} + \lambda_{K(k)} + \lambda_{IJ(ij)} + \lambda_{IK(ik)} + \lambda_{JK(jk)} + \lambda_{IJK(ijk)}.$$

En los estudios de sesgo interesa probar tres modelos, a partir de los cuales se pueden probar las hipótesis de interés sobre presencia y tipo de DIF (ver tabla 2). Si el único modelo que se ajusta es el saturado, que incluye todos los términos de interacción, se acepta que el ítem tiene DIF no uniforme; si el mejor modelo es el que excluye la interacción de tercer orden (ítem *x* grupo *x* magnitud del atributo), se acepta que el ítem presenta DIF uniforme; y si el mejor modelo excluye además la interacción entre grupo y respuesta al ítem, se afirmará que el ítem no presenta DIF. La tabla 2 sintetiza la relación entre el modelo elegido y la hipótesis sobre DIF. Siguiendo un algoritmo elegido[2] se realizan las estimaciones de los parámetros con la restricción de que la suma de todos los valores λ de un mismo efecto sea nula (igual a cero). Se trata de un análisis jerárquico, que, gracias a que los modelos son anidados, empieza con el modelo saturado (que se ajusta perfectamente a los datos) y va excluyendo términos no significativos empezando por los de orden superior, hasta llegar al modelo de efectos principales. En cada etapa del análisis se evalúa el ajuste del modelo que se basa en las diferencias existentes entre las frecuencias observadas y las esperadas a partir del mismo; y continúa comparando el ajuste del modelo anterior con el actual, habiendo excluido algún término. Según Bishop, Fienberg & Holland (1975) los dos estadísticos más frecuentemente usados para evaluar el ajuste son el χ^2 de Pearson y la razón de verosimilitud, conocida como G^2, los cuales siguen una distribución χ^2 con los mismos grados de libertad del modelo en cuestión. Dependiendo de cuál modelo resulte con mejor ajuste a los datos, se puede identificar los ítems DIF y el tipo de DIF, como se ilustra en la tabla 2.

[2] Según Fidalgo (1996a) los algoritmos más utilizados en los estudios sobre DIF son el de Newton-Rapson y el de ajuste proporcional iterativo.

Los modelos logit pueden verse como un caso especial de los modelos log-lineales, cuando la variable respuesta al ítem se considera dependiente de las otras dos y se realiza una transformación logit $log\,it(\,p\,) = ln\left(\frac{p}{1-p}\right)$ que no es otra cosa que el logaritmo del *odds* de los aciertos en el ítem, donde *p* es la proporción de acierto en el mismo. De esta manera, los tres modelos mencionados para verificar la existencia de DIF pueden expresarse en términos de la trasformación logit, como aparece en la tabla 2, donde los τ representan los diferentes efectos de acuerdo con la notación utilizada hasta ahora. El análisis sigue la misma lógica ya expuesta en relación con los modelos log-lineales; es decir, se trata de encontrar el modelo que mejor se ajuste a los datos y de acuerdo con él, concluir sobre la presencia o tipo de DIF del ítem analizado.

Tabla 2
Relación entre el modelo log-lineal y logit ajustado y la hipótesis sobre existencia y tipo de DIF

Hipótesis sobre DIF	Modelo log-lineal	Modelo logit
DIF no uniforme	$ln(e_{ijk}) = \lambda + \lambda_{I(i)} + \lambda_{J(j)} + \lambda_{K(k)} + \lambda_{IJ(ij)} + \lambda_{IK(ik)} + \lambda_{JK(jk)} + \lambda_{IJK(ijk)}$	$ln\left(p_{ijk} \big/ 1-p_{ijk}\right) = \tau + \tau_{J(j)} + \tau_{K(k)} + \tau_{JK(jk)}$
DIF uniforme	$ln(e_{ijk}) = \lambda + \lambda_{I(i)} + \lambda_{J(j)} + \lambda_{K(k)} + \lambda_{IJ(ij)} + \lambda_{IK(ik)} + \lambda_{JK(jk)}$	$ln\left(p_{ijk} \big/ 1-p_{ijk}\right) = \tau + \tau_{J(j)} + \tau_{K(k)}$
Ausencia de DIF	$ln(e_{ijk}) = \lambda + \lambda_{I(i)} + \lambda_{J(j)} + \lambda_{K(k)} + \lambda_{IK(ik)} + \lambda_{JK(jk)}$	$ln\left(p_{ijk} \big/ 1-p_{ijk}\right) = \tau + \tau_{K(k)}$

La aplicación de estos modelos a los análisis de sesgo de los ítems tiene la misma limitación ya comentada de los métodos que utilizan la puntuación observada en la prueba como medida de la magnitud del atributo. Siguiendo los mismos principios del procedimiento de dos etapas explicado antes para mejorar la eficacia del MH, Van Der Flier, Mellenbergh, Adèr & Wijn (1984) propusieron un procedimiento logit iterativo para purificar la medida de magnitud del atributo. El estudio de Kok, Mellenbergh & Van Der Flier (1985) y, más recientemente, el de Navas-Ara & Gómez Benito (2002) mostraron que purificar la escala de medida de la magnitud del atributo mejora la precisión del método. Sin embargo, Mellenbergh (1989) encontró que este procedimiento no es muy preciso cuando el porcentaje de ítem DIF en la prueba es grande, ya que la estimación de la magnitud del atributo debe hacerse solamente con los ítems sin DIF y no resulta muy precisa. De otra parte, Fidalgo & Paz Caballero (1995) en un estudio con datos simulados utilizando modelos log-lineales encontraron que este método tiene una tasa de detección bastante alta cuando la magnitud de DIF es grande, incluso moderada, y los grupos tienen tamaño superior a 200; sin embargo, comparado con otros métodos, como el MH, su error tipo I resulta demasiado alto; los autores concluyen que, aunque estos modelos presentan ventajas demostradas teóricamente por la

posibilidad de evaluar las relaciones entre todas las variables y de detectar tanto DIF uniforme y no uniforme, su uso no parece más aconsejable que el del MH en los estudios aplicados.

Regresión logística

La RL, procedimiento propuesto para la detección de DIF uniforme y no uniforme por Spray & Carlson (1986), Bennet, Rock & Kaplan (1987) y Swaminathan & Rogers (1990), puede verse como un caso especial de los análisis de regresión múltiple en el cual la variable dependiente es dicótoma; o como una extensión del MH donde la magnitud del atributo no se categoriza[3]. Utilizando una transformación logit, un modelo lineal para explicar la probabilidad de acierto en el ítem es $ln\left(\frac{p}{1-p}\right) = \tau_0 + \tau_1\theta + \tau_2 J + \tau_3(\theta J)$, que suele expresarse en términos de la conocida función logística como

$$p_i = \frac{e^{\tau_0+\tau_1\theta+\tau_2 J+\tau_3(\theta J)}}{1+e^{\tau_0+\tau_1\theta+\tau_2 J+\tau_3(\theta J)}},$$

donde: p_i es la probabilidad de acierto en el ítem y toma valores entre *0* y *1*
θ es la magnitud del atributo medido y entra en el modelo como una variable numérica medida a partir del puntaje total en la prueba
$J = 0$ ó *1* dependiendo del grupo al que pertenezca el examinado, y
τ_0, τ_1, τ_2 y τ_3 son los parámetros constante, del efecto de habilidad, del efecto de grupo y del efecto de interacción, respectivamente.

El análisis parte entonces de un modelo logístico que explica la probabilidad de acierto en el ítem en función de la magnitud del atributo, el grupo de pertenencia del examinado y la interacción entre éstos. Las estimaciones de dichos parámetros ($\hat{\tau}_0, \hat{\tau}_1, \hat{\tau}_2$ y $\hat{\tau}_3$) se obtienen por máxima verosimilitud. Si $\tau_3 = 0$ con $\tau_2 \neq 0$, el ítem tiene DIF uniforme y si la variable grupo se codifica de manera que se asigna el valor *1* al grupo de referencia y *0* al grupo focal, $\tau_2 > 0$ indica DIF a favor del grupo de referencia y $\tau_2 < 0$ indica DIF a favor del grupo focal. El DIF no uniforme ocurre cuando $\tau_3 \neq 0$; si $\tau_3 > 0$, el ítem favorece a los miembros del grupo de referencia de alta magnitud del atributo y a los miembros del grupo focal con baja magnitud del atributo, mientras que $\tau_3 < 0$ indica DIF a favor de las personas de baja magnitud del atributo del grupo de referencia y alta magnitud del atributo del grupo focal. Hidalgo Montesinos & Gómez Benito (2003) distinguen además DIF no uniforme simétrico y

[3] En Kleinbaum (1994), Bishop, Fienberg & Holland (1975) y Christensen (1997) se encuentran explicaciones detalladas sobre este tipo de modelos estadísticos y sus fundamentos. Aquí se presenta su aplicación a la detección de DIF

no simétrico; si $\tau_3 \neq 0$ y también $\tau_2 \neq 0$, el ítem muestra DIF no uniforme asimétrico; y si $\tau_2 = 0$, entonces se trata de DIF no uniforme simétrico. Estas hipótesis pueden someterse a prueba comparando la razón de verosimilitud, de los modelos de regresión incluyendo y sin incluir el parámetro de interés (ver Bishop, Fienberg & Holland, 1975). La diferencia en el logaritmo de la función de verosimilitud obtenida incluyendo y sin incluir el parámetro τ_3 constituye la prueba de DIF no uniforme; así mismo tal diferencia entre el modelo que incluye y excluye τ_2 sin τ_3, es la prueba de DIF uniforme.

La hipótesis de ausencia de DIF tanto uniforme como no uniforme, formulada en términos de los parámetros del modelo, es $H_0 : \tau_2 = \tau_3 = 0$, lo que indicaría que la probabilidad de acierto depende solamente de la magnitud del atributo medido. La diferencia entre el logaritmo de la función de verosimilitud entre el modelo compacto que incluye solamente el parámetro constante y el de magnitud del atributo $\left(p_i = \frac{e^{\tau_0 + \tau_1 \theta}}{1 + e^{\tau_0 + \tau_1 \theta}} \right)$ y el modelo saturado, que incluye todos los parámetros, sigue una distribución χ^2 con dos grados de libertad. De esta manera se puede someter a prueba la hipótesis de ausencia de ambos tipos de DIF simultáneamente. Sin embargo, Swaminathan & Rogers (1990) propusieron que la hipótesis de ausencia de DIF, tanto uniforme como no uniforme, puede formularse como $H_0 : C\hat{\tau}' = 0$ y someterse a prueba mediante el estadístico χ^2 con *2* grados de libertad

$$\chi^2_{RL} = \hat{\tau}' C' \left(C \sum C \right)^{-1} C \hat{\tau}'$$

donde: $\hat{\tau} = \left[\hat{\tau}_0 \quad \hat{\tau}_1 \quad \hat{\tau}_2 \quad \hat{\tau}_3 \right]$ representa el vector de estimaciones de los parámetros,

Σ es la matriz de varianzas y covarianzas entre las estimaciones de los parámetros, y

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Resulta evidente la similitud entre este modelo y los modelos logit comentados antes; puede mostrarse que las dos aproximaciones son equivalentes cuando en el modelo de regresión las variables independientes se codifican como variables Dummy. En las aplicaciones frecuentes de detección de DIF, la principal diferencia entre los dos métodos es que en los modelos logit, como en los métodos presentados antes, la magnitud del atributo se maneja como una variable categórica, mientras que en la RL se considera la naturaleza continua de la magnitud del atributo, característica que se menciona con frecuencia como una de las ventajas de la RL. French & Miller (1996), Zumbo, (1999) e Hidalgo Montesinos & Gómez Benito (2003) destacan además su facilidad y versatilidad para ajustarse al análisis de ítems politómicos o con más de dos grupos, mientras que Swaminathan & Rogers (1990), Millsap & Everson (1993) y Jodoin & Huff (2001) enfatizan en la capacidad de la RL para detectar tanto DIF uniforme como no uniforme, este último se evalúa introduciendo en el modelo un término de interacción entre la magnitud del atributo y el grupo. Swaminathan &

Rogers (1990) y Rogers & Swaminathan (1993) advierten, sin embargo, que el grado de libertad que se pierde al incluir un parámetro para detectar DIF no uniforme puede disminuir su poder para detectar DIF uniforme. Efectivamente, ajustando modelos con el parámetro de interacción, Herrera & Gómez Benito (en revisión) encontraron tasas de detección bastante bajas para la RL cuando se trató de detectar DIF uniforme, éstas apenas alcanzaron un nivel moderado (40%) con grupos iguales de 1500 examinados; sin embargo, para DIF no uniforme hallaron tasas satisfactoriamente altas con 500 examinados por grupo y tasas moderadas con el mismo tamaño del grupo de referencia y grupos focales entre 100 y 250.

A pesar de sus bondades, la RL ha recibido algunas críticas por el efecto que pueden tener los ítems DIF en el criterio de equiparación de los grupos, la inflación de su error tipo I (sobre todo con altos tamaños de muestra) y la carencia de alguna medida de la magnitud del DIF. Con respecto a la primera dificultad, compartida con todos los métodos que utilizan como criterio de equiparación la puntuación observada en la prueba, se han propuesto procedimientos iterativos, como los ya descritos, que buscan corregir el efecto de los ítems con DIF sobre los puntajes de los examinados en la prueba. En general se trata de identificar los ítems DIF y excluirlos del criterio de equiparación para romper la circularidad que se genera cuando éstos tienen algún peso en el criterio. Como se ha reportado con otros métodos, múltiples trabajos como los de Rogers & Swaminathan (1993), Gómez Benito & Navas-Ara, (1996) y Navas-Ara & Gómez Benito (2002), entre otros, han mostrado que el procedimiento iterativo de purificación del criterio de equiparación mejora el poder de la RL. Además, Hidalgo Montesinos & Gómez Benito (1996, 2003) y Gómez Benito & Hidalgo Montesinos (1997a) han utilizado también la purificación del criterio con regresión logística multinomial con resultados satisfactorios.

Con respecto a la carencia de una medida de la magnitud de DIF, Zumbo & Thomas (1996) y Zumbo (1999) propusieron una medida de mínimos cuadrados ponderados, denominada $R^2\Delta$ que permite evaluar la contribución de cada una de las variables en el modelo y entonces puede interpretarse en términos de la magnitud de DIF tanto uniforme como no uniforme. Mediante un estudio con datos simulados Jodoin & Gierl (2001) propusieron una clasificación de la magnitud de DIF, con base en los valores del $R^2\Delta$, así: DIF despreciable si $R^2\Delta < .035$, DIF moderado si se rechaza la hipótesis y $.035 < R^2\Delta < .07$ y, DIF amplio si se rechaza la hipótesis y $R^2\Delta > .07$. De acuerdo con los autores, esta categorización concuerda con otras como la de Zieky, (1993), adoptada por el ETS para el MH. Jodoin & Gierl (2001) y Jodoin & Huff (2001) evaluaron el funcionamiento de esta medida del efecto bajo diferentes condiciones de tamaño de muestra y distribución de la magnitud del atributo y encontraron que este criterio de clasificación no solamente provee una medida satisfactoria de la magnitud de DIF, sino que mantiene controlado el error tipo I, incluso con muestras grandes. Sin embargo, Hidalgo Montesinos & López Pina (2004) encontraron que tanto el sistema de clasificación de Zumbo & Thomas (1996) como el de Jodoin & Gierl (2001) fueron poco sensitivos, puesto que sólo clasificaron un máximo del 20% de ítems con DIF en las categorías de moderado o alto. Además, en el estudio de Jodoin & Gierl (2001) se encontraron tasas de detección bajas para DIF uniforme y moderadas para

DIF no uniforme con grupos de 250 examinados, las primeras solamente mejoraron hasta un nivel aceptable con grupos iguales de 1000 examinados por grupo.

Conclusiones

Este trabajo presentó una revisión de las técnicas más populares propuestas hasta el momento, con base en el análisis de tablas de contingencia, para detectar DIF en ítems dicótomos de pruebas psicológicas. Comparadas con otras técnicas como las que se basan en la Teoría de Respuesta al Ítem, las presentadas aquí resultan más sencillas y computacionalmente económicas, por lo que representan opciones viables para los constructores de pruebas interesados en introducir este tipo de análisis dentro de los procesos de evaluación y control de calidad de los instrumentos, disponiendo de muestras pequeñas y pocos argumentos para sustentar el cumplimiento de supuestos. Se propuso una clasificación de dichas técnicas en dos categorías: las que comparan proporciones de algunas celdas de la tabla y las que ajustan modelos para explicar las relaciones entre las variables de la misma. En la primera se incluyeron las aplicaciones del Ji cuadrado, el procedimiento de estandarización y el Mantel-Haenszel, mientras que en la segunda categoría se describieron el uso de los modelos log-lineales y logit y la regresión logística.

Dentro de la primera categoría, el procedimiento más popular es el Mantel-Haenszel; aunque, siguiendo las estrategias usadas por el ETS, una opción muy atractiva para los usuarios de los procedimientos puede ser la combinación del χ^2_{MH}, para probar la hipótesis de existencia de DIF, y la diferencia de proporciones estandarizada (DPE) en su métrica original o su transformación delta (DPE D-DIF), para describirlo. Disponer de una estrategia que permita clasificar la magnitud de DIF siguiendo algún criterio empírico como el propuesto por Dorans & Holland (1993) ayuda a ponderar el resultado estadístico en términos de su significación práctica. Una diferencia estadísticamente significativa con una magnitud de DIF que pueda describirse como despreciable, junto con el juicio y conocimiento del constructor del ítem, puede conducir al no rechazo de algunos elementos de la prueba que resultarían rechazados utilizando solamente el primer criterio. Aceptar un ítem con DIF puede tener implicaciones técnicas, éticas y sociales que han sido ampliamente discutidas, pero rechazar uno sin DIF eleva innecesariamente los costos en la construcción de instrumentos, costos que pueden ser importantes dependiendo del tipo de instrumento del que se trate. Jodoin & Huff (2001) hacen notar que este asunto cobra mayor importancia con la introducción de tecnologías como video, efectos de sonido y otras que buscan simular situaciones reales en las pruebas.

Un tema que ha despertado mucho interés es la identificación del tamaño mínimo necesario de los grupos para tener resultados satisfactorios. Se afirma frecuentemente que el MH, a diferencia de otros métodos como los que proponen modelos o los basados en la TRI, tiene un bajo requerimiento en tamaños de muestra; aunque los resultados de los múltiples trabajos al respecto apoyan esa ventaja comparativa, es necesario que los usuarios del método tengan un tamaño mínimo necesario y estén informados de las consecuencias de usar el método en otras circunstancias. Aunque la discusión no está cerrada y al evaluar la eficacia del

procedimiento deben considerar otros factores, la revisión de la literatura parece indicar que un tamaño mínimo para tener un poder aceptable en la detección de DIF uniforme y un adecuado control del error tipo I puede estar alrededor de los 200 examinados por grupo o, un mínimo de 125 en un grupo si el otro tiene al menos 500 examinados. Cuando los grupos estén entre 50 y 200, el usuario puede seguir la recomendación de Fidalgo, Ferreres & Muñiz (2004) y elevar el nivel de significancia hasta $\alpha \leq .2$, pero debe saber que en esas condiciones tendrá tasas de detección moderadas y el consecuente aumento en la probabilidad de detectar como DIF ítems que no lo son (error tipo I). De otra parte, aunque el DIF no uniforme es menos frecuente en la práctica, debe tenerse en mente que el MH ha mostrado limitaciones en su detección y que sólo se tendrán tasas satisfactorias con grupos muy grandes (1500 examinados) sobre todo en con dificultad media.

Dentro de los métodos de la segunda categoría, los que ajustan modelos para explicar las relaciones entre las variables involucradas en la tabla de contingencia, tal vez el que ha suscitado mayor interés es la aplicación de la regresión logística. Este procedimiento tiene algunas ventajas comparativas: mejores tasas de detección de DIF no uniforme que el MH, facilidades de aplicación cuando se tienen más de dos grupos o ítems politómicos y menos costo computacional que las técnicas basadas en la TRI. Aunque la producción de trabajos sobre los tamaños de muestra es menor que la del MH, los resultados de Jodoin & Huff (2001) no son muy satisfactorios con grupos iguales de 250, los de Herrera & Gómez Benito (en revisión) parecen indicar que con 500 examinados por grupo se pueden obtener resultados satisfactorios para DIF no uniforme, otros trabajos reportan mejores resultados con grupos de mayor tamaño y es de esperarse que los requerimientos de muestra para estimar los parámetros del modelo sean mayores que los del MH. De otra parte, utilizando el procedimiento estándar, la RL muestra bajo poder para detectar DIF uniforme; sin embargo, aún falta mayor investigación sobre el comportamiento de las propuestas de Zumbo & Thomas (1997) y Jodoin & Huff (2001).

Finalmente vale la pena enfatizar que aunque las técnicas para detectar DIF constituyen un avance importante en la construcción de instrumentos de medida de condiciones óptimas, el uso de tales procedimientos es una tarea necesaria pero no suficiente para el análisis de sesgo. La detección de DIF es sólo la evidencia empírica de que el ítem tiene comportamiento diferente en grupos diferentes igualados por magnitud del atributo con sustento en la estadística, pero el análisis de sesgo pone en juego el juicio del constructor del instrumento e implica un profundo conocimiento de las teorías acerca de lo que mide el instrumento y sus posibles relaciones con otras variables como género, cultura, etc. Una tarea que apenas está iniciada y que ayudaría en este complejo proceso es la identificación de las razones para que ocurra el DIF. Hasta el momento la teoría más aceptada es la de la multidimensionalidad propuesta por Ackerman (1992) pero recientemente Borsboom, Mellenbergh & van Heerden (2002) han encontrado que ésta no parece ajustarse a medidas que pueden necesitar escalas relativas a un grupo, como ocurre con ítems de pruebas como las de personalidad o intereses.

Referencias

Ackerman, T. (1992). A Didactic Explanation of Item Bias, Item Impact, and Item Validity From a Multidimensional Perspective. *Journal of Educational Measurement. 29*(1), 67-91.

Angoff, W. H. (1993). Perspectives on Differential Item Functioning Methodology. En P.W.Holland & H. Wainer (Eds.), *Differential Item Functioning*, New Jersey: Lawrence Erlbaum Associates, Inc.

Baker, F. B. (1981). A Criticism of Scheuneman´s Item Bias Techinique. *Journal of Educational Measurement. 18*(1), 59-62.

Bennet, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT differential item performance for nine handicapped group. *Journal of Educational Measurement. 24*, 41-55.

Binet, A. & Simon, T. (1916). *The development of intelligence in children*. New York: Amo.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge: MIT Press.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (Ed) (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement. 26* (4), 433-450.

Camilli, G. (1993). The Case against Item Bias Detection Techniques Based on Internal Criteria: Do Item Bias Procedures Obscure Test Fairness Issues? En P.W.Holland & H. Wainer (Eds.), *Differential Item Functioning*, pp. 397-417. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Camilli, G. & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. 4. United States: SAGE Publications.

Christensen, R. (1997). *Log-Linear Models and Logistic regression*. (2 ed.) New York: Springe.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effect of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education. 6*(4), 269-279.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo Study of Factor That affect the Mantel-Haenszel and Standardization Measures of Differential Item Functioning. En P.W.Holland & H. Wainer (Eds.), *Differential Item Functioning*, pp. 137-166. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publisher.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education. 2*, 217-233.

Dorans, N. J. & Holland, P. W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. En P.W.Holland & H. Wainer (Eds.), *Differential Item Functioning*, New Jersey: Lawrence Erlbaum Associates, Inc.

Dorans, N. J. & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach*. (Research Rep. No 83-9). Princeton, NJ: Educational Testing Service.

Dorans, N. J. & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on Scholastic Aptitude Test. *Journal of Educational Measurement. 23*, 355-368.

Eelles, K., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. (1951). *Intelligence and cultural differences*. Chicago: University of Chicago Press.

Ferreres, T. (1998). *Funcionamiento diferencial de los items de una prueba de aptitud intelectual en función de la lengua familiar y la lengua de escolarización*. Tesis doctoral inédita. Valencia: Universidad de Valencia.

Fidalgo, Á. M. (1996a). Funcionamiento Diferencial de los Items. En J.Muñiz (Ed.), *Psicometría*, pp. 371-455. Madrid: Editorial Universitas, S.A.

Fidalgo, Á. M. (1996b). *Funcionamiento diferencial de los ítems. Procedimiento Mantel-Haenszel y modelos loglineales.* Tesis doctoral no publicada. Oviedo: Universidad de Oviedo.

Fidalgo, Á. M., Ferreres, D., & Muñiz, J. (2004). Utility of the Mantel-Haenszel ´procedure for detecting differential item functioning in small samples. *Educational and Psychological Measurement. 64*(6), 925-936.

Fidalgo, Á. M. & Mellenbergh, G. J. (1995). *Evaluación del procedimiento Mantel-Haenszel frente al método logit iterativo en la detección del funcionamiento diferencial de los ítems uniforme y no uniforme.* Comunicación presentada al IV Simposio de Metodología de las Ciencias del Comportamiento. La Manga del Mar Menor.

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

Fidalgo, Á. M. & Paz Caballero, M. D. (1995). Modelos lineales logarítmicos y funcionamiento diferencial de los ítems. *Anuario de Psicología. 1995*(64), 57-66.

French, A. W. & Miller, T. (1996). Logistic Regression and its Use in Detecting Differential Item Functioning in Polytomous Items. *Journal of Educational Measurement. 33*(3), 315-333.

Gómez Benito, J. & Hidalgo Montesinos, M. D. (1997a). *A Comparison of Two Procedures of Ability Purification on the Detection of Differential Item Functioning Using Multinomial Logistic Regression*. Poster presented at the 10th European Meeting of the Psychometric Society. Santiago de Compostela, Spain.

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

Gómez Benito, J., Hidalgo Montesinos, M. D., Guilera Ferré, G., & Moreno Torrente, M. (2005). A bibliometric study of differential item functioning. *Scientometrics. 64*(1), 3-16.

Gómez Benito, J. & Navas-Ara, M. J. (1996). Detección del funcionamiento diferencial de los ítems mediante regresión logística: Purificación paso a paso de la habilidad. *Psicológica. 17*, 397-411.

Gómez Benito, J. & Navas-Ara, M. J. (2000). A Comparison of $X^{2,\ RFA}$ and IRT Based Procedures in the Detection of DIF. *Quality & Quantity. 34*, 17-31.

Gómez-Benito, J. & Hidalgo-Montesinos, M. D. (2003). Desarrollos recientes en psicometría. *Avances en Medición. 1*, 17-36.

Gottfredson, L. S. (Ed) (1997). Intelligence and social policy. *Intelligence. 24* (Special issue), 1-320.

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. (1993). Advances in the Detection of Differentially Functioning Test Items. *European Journal of Psychological Assessment. 9*(1), 1-18.

Herrera, A. N. & Gómez Benito, J. (en revisión). *The Effect of Sample Size Ratio on the Power and Type I Error of the Mantel-Haenszel and Logistic Regression Procedures in the Detection of Differential Item Functioning*. Manuscript submitted for publishing.

Herrera, A. N., Sánchez Pedraza, R., & Gómez Benito, J. (2001). Funcionamiento diferencial de los Items: Una revisión conceptual y metodológica. *Acta Colombiana de Psicología. 5*, 41-61.

Hidalgo Montesinos, M. D. & Gómez Benito, J. (1996). *The Effect of Ability Purification on the Evaluation of Differential Item Funtioning with the Technique of Multinomial Logistic Regression*. Paper presented at the 20th biennial conference of the Society for Multivariate Analysis in the Behavioral Sciences. Barcelona, Spain.

Hidalgo Montesinos, M. D. & Gómez Benito, J. (2003). Test Purification and the Evaluation of Differential Item Functioning with Multinomial Logistic Regression. *European Journal of Psychological Assessment. 19*(1), 1-11.

Hidalgo Montesinos, M. D. & López Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement. 64*(6), 903-915.

Hills, J. (1990). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice. 8*, 5-11.

Holland, P. W. & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty*. (Research report 85-43). Princeton, NJ: Educational Testing Service.

Holland, P. W. & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure*. (Tecnical report N° 86-89). Princeton, NJ: Educational Testing Service.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and Mantel-Haenszel procedure. En H.Wainer & H. I. Braun (Eds.), *Test Validity*, pp. 129-145. Hillsdale, N.J.: Erlbaum.

Holland, P. W. (1989). A note on the covariance of the Mantel-Haenszel log-odds estimator and the sample marginal rates. *Biometrics. 45*, 1009-1015.

Ironson, G. (1982). Use of Chi-square and Latent Trait Approaches for Detecting Item Bias. En R.Berck (Ed.), *Handbook of Methods for Detecting Test Bias*, pp. 117-160. Baltimore: The Jons Hopkins University Press.

Jensen, A. R. (1969). How much can we boast IQ and scholastic achievement? *Harvard Educational Review. 39*, 1-123.

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

Jensen, A. R. (1998). The *g* factor in the design of education. En R.J.Sternberg & W. M. Williams (Eds.), *Intelligence, instruction, and assessment*, pp. 111-131. Mahwah, NJ: Erlbaum.

Jensen, A. R. (2000). TESTING. The Dilemma of Group Differences. *Psychology, Public Policy, and Law. 6*(1), 121-127.

Jodoin, M. G. & Gierl, M. J. (Ed) (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education. 14* (4), 329-349.

Jodoin, M. G. & Huff, K. L. (2001). *Examining type I error and power rates when ability distribution are unequal with the logistic regresion procedure for DIF detection*. Paper presented at Annual metting of the National Council on Measurement in Education. Seattle.

Kleinbaum, D. G. (1994). *Logistic regression. A self learning text*. New York: Springer.

Knoke, D. & Burke, P. J. (2000). *Log-linear models*. Beverly Hills: 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.

Kulick, E. & Dorans, N. J. (1983a). *Assessing the unexpected differential item functioning of oriental candidates on SAT form CSAG and TSWE Form E33*. (Statistical Rep. No 83-106). Princeton, NJ: Educational Testing Service.

Kulick, E. & Dorans, N. J. (1983b). *Assessing the unexpected differential item performance of candidates reporting different levels of father's education on SAT form CSA2 and TSWE Form E29*. (Statistical Rep. No 83-27). Princeton, NJ: Educational Testing Service.

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

Mazor, K. M., Clauser, B. E., & 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 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*, 105-118.

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

Miller, M. D. & Oshima, T. C. (1992). Effect of Sample Size, Number of Biased Items, and Magnitude of Bias on a Two-Stage Item Bias Estimation Method. *Applied Psychological Measurement. 16*(4), 381-388.

Millsap, R. & Everson, H. (1993). Methodology Review: Statistical Approaches for Assessing Measurement Bias. *Applied Psychological Measurement. 17*(4), 297-334.

Muñiz, J. (1998). *Teoría Clásica de los Test*. Madrid: Ediciones Pirámide.

Narayanan, P. & 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.

Narayanan, P. & Swaminathan, H. (1996). Identification of Items that Show Nonuniform DIF. *Applied Psychological Measurement. 20*, 257-274.

Navas-Ara, M. J. & Gómez Benito, J. (1994). *Comparison of several bias detection techniques.* (Paper presented at the 23rd. International Congress of Applied Psychology ed.) Madrid.

Navas-Ara, M. J. & Gómez Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment. 18*(1), 9-15.

Phillips, A. & Holland, P. W. (1987). Estimation of the variance of the Mantel-Haenszel log-odds ratio estimate. *Biometrics. 43*, 425-431.

Powers, D. A. & Xie, Y. (2000). *Statistical methods for categorical data analysis*. San Diego: Academic Press.

Robins, J., Breslow, N., & Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics. 42*, 311-323.

Rogers, H. J. & Swaminathan, H. (1993). A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement. 17*(2), 105-116.

Roussos, L. A. & Stout, W. F. (1996). Simulation Studies of the Effects of Small Sample Size an Studied Item Parameters on SIBTEST and Mantel-Haenszel Type I Error Performance. *Journal of Educational Measurement. 33*(2), 215-230.

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

Scheuneman, J. D. (1981). A response to Baker's criticism. *Journal of Educational Measurement. 18*, 63-66.

Shealy, R. & Stout, W. F. (1993). A Model Based Standardization Approach that Separates True Bias/DIF From Group Ability Differences and Detects Test Bias/DTF as well as Item Bias/DIF. *Psychometrika. 58*(2), 159-194.

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

Stern, W. (1914). *The psychological methods of testing intelligence*. Baltimore: Warwick y York.

Swaminathan, H. & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement. 27*(4), 361-370.

Uttaro, T. & Millsap, R. (1994). Factors Influencing the Mantel-Haenszel Procedure in the Detection of Differential Item Functioning. *Applied Psychological Measurement. 18*(1), 15-25.

Van Der Flier, H., Mellenbergh, G. J., Adèr, H. J., & Wijn, M. (1984). An iterative item bias detection methods. *Journal of Educational Measurement. 21*(2), 131-145.

Zieky, M. (1993). Practical Questions in the Use of DIF Statistics in Test Development. En P.W.Holland & H. Wainer (Eds.), *Differential Item Functioning*, pp. 337-347. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Otawa: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. & Thomas, D. R. (1996). *A measure of DIF effect size using logistic regression procedure*. (Paper presented at the National Board of Medical Examiners ed.) Philadelphia.

Zumbo, B. D. & Thomas, D. R. (1997). *A mesure of effect size for a model-based approach for studying DIF*. Working paper of the Edgeworth Laboratory for Quantitative Behavioral Science. Prince George, Canada: University of Northern British Columbia.

Manuscrito recibido en Enero de 2005
Aprobado para publicación en Marzo de 2005