Psicothema 2005. Vol. 17, nº 3, pp. 509-515
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG


# SOFTWARE, INSTRUMENTACIÓN Y METODOLOGÍA

## Regresión logística: alternativas de análisis en la detección del funcionamiento diferencial del ítem

M.ª Dolores Hidalgo Montesinos, Juana Gómez Benito* y José Luis Padilla García**
Universidad de Murcia, * Universidad de Barcelona y ** Universidad de Granada

El presente estudio compara varias estrategias analíticas para la detección del Funcionamiento Diferencial del Ítem (DIF) mediante regresión logística. Las distintas alternativas de análisis se comparan sometiendo a contrastación distintos tipos de ítems en los que se han simulado diferentes condiciones de DIF. En todos los casos se ha trabajado con tests de 75 ítems dicotómicos. Se simularon tres tipos de DIF (uniforme, no-uniforme simétrico y no-uniforme asimétrico) con diferentes condiciones de cantidad de DIF (0.0, 0.3 y 1.0 para el parámetro de dificultad, y 0.0, 0.25 y 1.0 para el parámetro de discriminación). En términos generales, si atendemos tanto al criterio de mayor potencia en la detección del DIF como de menor costo computacional se recomienda el uso del análisis de regresión logística implementando la estrategia que somete a comprobación la presencia conjunta de DIF uniforme y no-uniforme, complementándola con la significación estadística de la interacción para distinguir entre los dos tipos de DIF.

*Logistic regression: analytic strategies in differential item functioning detection.* The present study compares several analytic strategies, whose relative efficacy has yet to be evaluated, for detecting Differential Item Functioning (DIF) by means of logistic regression. The strategies are compared by checking different item types in which various DIF conditions are simulated. In all cases 75-item, dichotomous response tests were used. Three types of DIF (uniform, symmetric non-uniform and asymmetric non-uniform) were simulated with three conditions for the amount of DIF (0.0, 0.3 and 1.0 for the difficulty parameter, and 0.0, 0.25 and 1.0 for the discrimination parameter). In general, and according to the criteria of greatest power in detecting DIF and low computational cost, it is recommended that applied psychologists and educators who analyse, translate and adapt tests, use logistic regression analysis with the strategy that checks for the presence of both uniform and non-uniform DIF; this should be complemented by calculating interaction significance in order to distinguish between the two types of DIF.

Uno de los indicadores de calidad de los instrumentos de medida psicológica es que presenten las mismas propiedades psicométricas, independientemente de los grupos que se están sometiendo a evaluación; en otras palabras, que sus características métricas sean invariantes a través de los distintos grupos sobre cuya ejecución en el test o cuestionario se pretenden hacer interpretaciones comparativas. En este contexto, se dice que un ítem muestra «Funcionamiento Diferencial del Ítem» (Differential Item Functioning, DIF) cuando presenta propiedades psicométricas diferentes en función del grupo en el que ha sido administrado, después que los grupos han sido igualados en el rasgo o habilidad medida por el test (Angoff, 1993). Así, en el caso de los ítems de rendimiento o aptitud se considera que un ítem presenta DIF cuando la probabilidad de obtener una respuesta correcta es diferente para grupos de personas igualados en la variable medida por el test.

Normalmente los grupos bajo estudio suelen ser dos, que tradicionalmente se denominan focal y referencia, para hacer la distinción entre el grupo cuyas respuestas han sido utilizadas para analizar el funcionamiento del test durante su construcción (grupo de referencia) y el grupo donde el test se pretende aplicar, pero para el que se sospecha que las propiedades psicométricas de los ítems puedan tener valores distintos (grupo focal). Si bien, cuando un ítem presenta DIF en la mayoría de las ocasiones suele ser unifor-

Fecha recepción: 6-4-04 • Fecha aceptación: 22-12-04
Correspondencia: M.ª Dolores Hidalgo Montesinos
Facultad de Psicología
Universidad de Murcia
30080 Murcia (Spain)
E-mail: mdhidalg@um.es

me, es decir, la diferencia en la probabilidad de acertar el ítem siempre favorece a uno de los grupos a lo largo del continuo de habilidad medido por el test, también se han identificado otros tipos: DIF no-uniforme simétrico y DIF no-uniforme asimétrico (Mellenbergh, 1982; De Ayala, Kim, Stapleton y Dayton, 1999). En estos últimos, la diferencia entre la probabilidad de dar una respuesta correcta al ítem del grupo de referencia y la del grupo focal, no es constante a lo largo del continuo de habilidad.

El análisis de regresión logística (Swaminathan y Rogers, 1990; Rogers y Swaminathan, 1993) es uno de los métodos de comprobada eficacia para su uso en la detección tanto del DIF uniforme como no-uniforme (Clauser y Mazor, 1998). Trabajos previos en este campo han demostrado que este procedimiento produce resultados similares a los obtenidos con el estadístico Mantel-Haenszel cuando se trata de detectar DIF uniforme y una mayor potencia estadística para detectar ítems con DIF no-uniforme (Clauser, Nungester, Mazor y Ripkey, 1996; Ferreres, Fidalgo y Muñiz, 2000; Hidalgo y Gómez, 2000; Hidalgo y López, 2004; Narayanan y Swaminathan, 1996; Rogers y Swaminathan, 1993).

Swaminathan y Rogers (1990) y Rogers y Swaminathan (1993) propusieron el análisis de regresión logística para la detección del DIF uniforme y no-uniforme en ítems dicotómicos. Esta utilización de la regresión logística se basa en el modelado estadístico de la probabilidad de obtener una respuesta correcta al ítem que se considera función de dos variables: la pertenencia al grupo (referencia o focal) y el nivel de habilidad de los sujetos (puntuación empírica u observada en el test o bien el nivel de habilidad estimado bajo algún modelo de respuesta al ítem). La puntuación total observada en el test se utiliza para igualar a los sujetos respecto de la habilidad medida por el test y, a diferencia de otros procedimientos como los modelos loglineales o el estadístico Mantel-Haenszel, el análisis de regresión logística trata dicha puntuación total de forma continua. El modelo estadístico de regresión logística para la detección del DIF considera la respuesta al ítem como la variable dependiente, mientras que las variables explicativas son la puntuación observada del sujeto en el test, la pertenencia al grupo y la interacción entre puntuación observada y pertenencia a grupo. El efecto de las variables explicativas sobre la variable dependiente puede evaluarse utilizando distintas pruebas de significación y estrategias analíticas (Clauser y Mazor, 1998; Gómez e Hidalgo, 1997; Hosmer y Lemeshow, 1989; Rogers y Swaminathan, 1993).

El presente estudio compara el comportamiento de tres estrategias analíticas posibles cuando se utiliza el análisis de regresión logística para la detección del DIF en ítems dicotómicos, cuya eficacia relativa aún no ha sido evaluada. Se presenta un estudio específico que pone a prueba la capacidad de detección del DIF de cada estrategia y su nivel de errores de clasificación en distintas condiciones de DIF.

## Análisis de regresión logística

La ecuación general para el análisis de regresión logística adopta la forma:

$$P(y=1/X)=\frac{e^{z}}{1+e^{z}}$$

donde *P(y= 1/X)* es la probabilidad de obtener una respuesta correcta condicionado a X (puntuación observada del sujeto en el test) y *z* representa la combinación lineal de las variables predictoras. En el análisis del DIF, el modelo de regresión logística se parametriza en los siguientes términos:

$$z=\beta_0+\beta_1X+\beta_2G+\beta_3XG$$

donde *X* es la puntuación observada de un sujeto en un test y *G* la variable de grupo de pertenencia de los sujetos (cultura, idioma, raza, sexo, edad,…); además, $\beta_0$ es la intercepción, $\beta_1$ es el coeficiente para la habilidad o puntuación total observada en el test, $\beta_2$ es el coeficiente para la variable grupo de pertenencia (referencia o focal), y $\beta_3$ es la interacción entre la puntuación observada en el test y el grupo.

Bajo esta formulación, un ítem muestra DIF uniforme si el efecto del grupo (*G*) resulta estadísticamente significativo, mientras que la interacción habilidad por grupo (*XG*) no ejerce ningún efecto sobre el ítem. Por el contrario, si la interacción *XG* resulta estadísticamente significativa, el ítem presentaría DIF no-uniforme.

La aplicación de la regresión logística para la detección del DIF puede realizarse con distintas estrategias. La primera estrategia se basa en la comparación de modelos anidados. Se ajustan tres modelos en distintas etapas. En la primera etapa, se ajusta el modelo base de ausencia de DIF (Modelo 1), donde se introduce en la ecuación la puntuación total del sujeto en el test (*X*). En la segunda etapa, se añade a la ecuación la variable de agrupamiento (*G*), ajustándose el modelo de DIF uniforme (Modelo 2). Por último, en la tercera etapa se introduce en la ecuación la interacción entre el grupo y la puntuación total en el test, valorándose el ajuste del modelo de DIF no-uniforme o modelo completo (Modelo 3). En el cuadro 1 se presentan los distintos modelos a ajustar en la detección y evaluación del DIF y su correspondiente parametrización.

En esta estrategia de análisis la evaluación del DIF se efectúa comprobando la significación del efecto de las sucesivas variables que se van introduciendo en el modelo. Así, mediante la comparación entre el valor de la razón de verosimilitud[1] del Modelo 1 (que expresa ausencia de DIF ya que la respuesta al ítem sólo depende del nivel de la habilidad del sujeto) con el del Modelo 2 (que introduce además el efecto de grupo), se obtiene una prueba para el DIF uniforme; el estadístico de comparación representa el cambio en el ajuste desde una ecuación a la otra, sigue una distribución $\chi^2$ con 1 grado de libertad, y se suele denominar $G^2$ de diferencia[2] (Bishop, Fienberg y Holland, 1975). Por su parte, si se compara el valor de verosimilitud del Modelo 2 con el del Modelo 3, se puede probar la existencia de DIF no uniforme; este estadístico de diferencia sigue también una distribución $\chi^2$ con 1 grado de libertad. Así pues, en este proceso de comparación de modelos lo que se evalúa es la mejora explicativa al introducir un nuevo término al modelo.

La segunda estrategia para la detección del DIF mediante el análisis de regresión logística consiste en realizar una prueba simultánea de la presencia de DIF uniforme y no-uniforme. Esta hipótesis conjunta se puede someter a comprobación comparando el

*Cuadro 1*
Modelos de regresión logística a analizar en el estudio del DIF

| | |
|---|---|
| Etapa 1 (Modelo 1) | $z=\beta_0+\beta_1X$ |
| Etapa 2 (Modelo 2) | $z=\beta_0+\beta_1X+\beta_2G$ |
| Etapa 3 (Modelo 3) | $z=\beta_0+\beta_1X+\beta_2G+\beta_3XG$ |

valor de verosimilitud del modelo sin DIF (Modelo 1) con el del modelo completo (Modelo 3); en este caso el estadístico $G^2$ de diferencia sigue una distribución $\chi^2$ con 2 grados de libertad. El uso de esta estrategia de análisis no permite evaluar el tipo de DIF que presenta el ítem.

Las dos estrategias anteriores obligan a ajustar más de un modelo de regresión logística. La tercera estrategia consiste en ajustar solamente el modelo completo, incluyendo todos los términos (*X, G* y *X G)*, y comprobar la significación de los coeficientes del modelo asociados a cada término, usando el estadístico de Wald[3] que, siempre que la variable de agrupamiento tenga dos niveles, se distribuye según una distribución $\chi^2$ con 1 grado de libertad (sólo bajo la hipótesis nula y en muestras grandes). Si únicamente el coeficiente de la variable grupo ($\beta_2$) es significativo, el ítem muestra DIF uniforme; si sólo es significativo el coeficiente de la interacción ($\beta_3$), el DIF detectado es no-uniforme simétrico; finalmente, si ambos coeficientes son significativos, el ítem está afectado por DIF no-uniforme asimétrico.

En el cuadro 2 se presenta un resumen de las tres estrategias de detección del DIF usando regresión logística.

## Método

En este estudio se utilizaron datos simulados para ejemplificar el funcionamiento de las distintas estrategias de detección del DIF con el análisis de regresión logística. El uso de datos simulados nos permite definir de partida diferentes condiciones experimentales de DIF y comprobar el efecto en la identificación del mismo usando como técnica de análisis la regresión logística.

### *Generación del DIF y factores manipulados*

Se simularon 10 tests de respuesta dicotómica, estando formados cada uno de ellos por 14 ítems con DIF y 61 ítems sin DIF; en total, cada test constaba de 75 ítems de los cuales aproximadamente un 20% de los ítems contienen DIF. Este porcentaje representa una situación adversa para la identificación del DIF, sin embargo, mientras que Narayanan y Swaminathan (1994) señalan que, por ejemplo, en tests de rendimiento no es habitual encontrar más de un 10% ó 15% de los ítems con DIF, Miller y Linn (1988) apuntan que también se encuentran tests con un 20% y un 40% de ítems sesgados. Además, Gierl, Gotzmann y Boughton (2004) señalan que en situaciones de traducción y adaptación de tests, el porcentaje de ítems con DIF en el test suele ser bastante elevado, superior al 20%. De todos modos, Narayanan y Swaminathan (1996) encontraron pocas diferencias (sólo un 4%) en las tasas de identificación correcta de DIF usando regresión logística en tests en los que existía DIF, respectivamente, en un 10% y 20% de los ítems.

Para el análisis del DIF, un total de 140 ítems con DIF, asignados aleatoriamente a cada uno de los 10 tests, fueron sometidos a estudio. Esos 140 ítems fueron elegidos variando sus niveles de dificultad y discriminación, de este modo se seleccionaron cinco niveles de dificultad para el grupo de referencia, representando respectivamente ítems muy fáciles, fáciles, de dificultad media, difíciles y muy difíciles. En concreto los valores para el parámetro de dificultad fueron de –1.5, -1.0, 0, 1, y 1.5. Además estos valores se cruzaron con cuatro niveles del parámetro de discriminación 0.25, 0.60, 0.90 y 1.25. Para los 140 ítems bajo estudio se simularon tres tipos de DIF: uniforme, no-uniforme simétrico y no-uniforme asimétrico. El DIF fue generado como diferencias entre los parámetros del ítem para el grupo de referencia y para el grupo focal. Los tres niveles de cantidad de DIF fueron para el parámetro de dificultad: 0.0, 0.3 y 1.0; y para el parámetro de discriminación: 0.0, 0.25 y 1.0. Se simuló el DIF uniforme variando el parámetro de dificultad para los dos grupos. El DIF no uniforme fue simulado variando los parámetros de dificultad y discriminación para los dos grupos (no-uniforme asimétrico), o manteniendo el mismo parámetro de dificultad y variando el parámetro de discriminación para los dos grupos (no-uniforme simétrico). El DIF se simuló siempre en el mismo sentido, favoreciendo al grupo de referencia.

En la tabla 1 aparece, para cada condición de DIF, una estimación del tamaño del DIF; este tamaño del efecto fue calculado usando las medidas de área propuestas por Raju (1988) que nos permite conocer el área entre las funciones de respuesta al ítem para los dos grupos. Las diferencias fueron generadas para representar distintos tamaños del DIF, desde un tamaño del efecto pequeño (0.3) a un tamaño del efecto alto (1.0 y 1.50) (Narayanan y Swaminathan, 1996).

La habilidad de los sujetos se generó aleatoriamente según una distribución normal tipificada N(0,1) para un intervalo de $\theta$ desde –3 a +3. Las investigaciones sobre la eficacia de la regresión logística en la detección del DIF señalan que, al igual que otras técnicas de análisis, la potencia de la misma se ve incrementada con el aumento del tamaño muestral (Narayanan y Swaminathan, 1996; Mazor, Clauser y Hambleton, 1992). En este trabajo, se ha utilizado un tamaño muestral grande de 1000 sujetos tanto para el grupo focal como para el grupo de referencia.

### *Generación de las matrices de respuestas a los ítems*

En la generación de las matrices de respuestas a los ítems se utilizó un programa informático construido para tal fin. Estas ma-

*Cuadro 2*
Estrategias de detección del DIF con regresión logística

| | | |
|---|---|---|
| Estrategia 1: | DIF uniforme | Comparación del modelo 1 con el 2 |
| | DIF no-uniforme | Comparación del modelo 2 con el 3 |
| Estrategia 2: | DIF uniforme y no-uniforme | Comparación del modelo 1 con el 3 |
| Estrategia 3: | DIF uniforme | Significación del coeficiente $\beta_2$ en el modelo 3 |
| | DIF no-uniforme | Significación del coeficiente $\beta_3$ en el modelo 3 |

*Tabla 1*
Magnitud del DIF generado

| a-dife | b-dife | a= .25 | .60 | .90 | 1.25 | Promedio |
|---|---|---|---|---|---|---|
| 1.0 | 1.0 | 2.74 | 1.22 | 1.04 | 1.00 | 1.50 |
| 1.0 | 0.0 | 2.61 | 0.85 | 0.48 | 0.29 | 1.00 |
| .25 | 0.3 | 1.65 | 0.48 | 0.33 | 0.30 | 0.69 |
| .25 | 0.0 | 1.63 | 0.40 | 0.20 | 0.11 | 0.58 |
| 0 | 1.0 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0 | 0.3 | 0.30 | 0.30 | 0.30 | 0.30 | 0.30 |

trices de respuestas fueron generadas bajo el modelo de TRI de tres parámetros (3-p) modificado, fijando a 0.20 el parámetro de pseudo-azar. El procedimiento de obtención de las respuestas fue del siguiente modo: partiendo de los parámetros verdaderos de discriminación, dificultad y pseudo-azar de los ítems y de los parámetros verdaderos de habilidad de los sujetos, se obtienen los distintos valores de probabilidad de acertar cada uno de los ítems del test en cada valor de habilidad. Cada valor de probabilidad se compara con un número pseudo aleatorio obtenido de una distribución uniforme con valores comprendidos entre 0 y 1. Si el valor de probabilidad obtenido por el sujeto es menor que el generado por la distribución uniforme, se considera que el sujeto ha fallado el ítem. Por el contrario, si el valor de probabilidad obtenido es mayor o igual al generado, se considera que el sujeto ha acertado el ítem y se le asigna un valor de 1.

*Análisis del DIF*

Cada uno de los 140 ítems fue analizado con el programa SPSS (v. 10) utilizando el procedimiento de análisis de regresión logística binaria ejecutándolo para cada una de las estrategias bajo estudio. Además, se utilizó como variable de equiparación (criterio de comparación de los grupos) la puntuación observada en el test.

Por último, en el análisis del DIF, independientemente de la estrategia utilizada, el nivel de confianza se estableció al 95%.

## Resultados

En general, el número de ítems con DIF correctamente identificados fue mayor cuando se utilizó la estrategia 2 que cuando se utilizaron la estrategia 1 o la 3.

En las tablas 2 a la 4 se presentan los resultados de identificación del DIF, detallados para cada uno de los 140 ítems del test, para cada una de las estrategias de análisis sometidas a estudio. Así, podemos observar que cuando las diferencias en el parámetro de discriminación son mayores el número de ítems correctamente identificado también es mayor. Del mismo modo, cuando las diferencias entre el grupo focal y el de referencia en el parámetro de dificultad fueron mayores, un mayor número de ítems con DIF fueron identificados correctamente (IC). De hecho, las tasas de IC fueron más altas cuando se dieron diferencias elevadas tanto en *a* (≥0.25) como en *b* (≥0.30).

Además, los resultados muestran que los ítems con DIF más discriminativos, en general, fueron ligeramente peor detectados que los ítems menos discriminativos, aunque este resultado parece depender también de la dificultad del ítem. Esto puede ser

*Tabla 2*
Ítems identificados como DIF para la estrategia 1

| | | b= -1.5 | | | | b= -1.0 | | | | b= 0.0 | | | | b= 1.0 | | | | b= 1.5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **a-dife** | **b-dife** | **a= .25** | **.60** | **.90** | **1.25** | **a= .25** | **.60** | **.90** | **1.25** | **a= .25** | **.60** | **.90** | **1.25** | **a= .25** | **.60** | **.90** | **1.25** | **a= .25** | **.60** | **.90** | **1.25** |
| 1.0 | 1.0 | / | X | X | * | X | * | * | * | X | * | * | X | * | X | X | X | X | X | X | X |
| 1.0 | 0.0 | X | X | X | * | X | X | X | X | X | / | / | X | X | X | / | | * | * | * | |
| .25 | 0.3 | X | / | | | X | | | | / | | * | * | * | * | * | | * | | * | |
| .25 | 0.0 | X | X | * | * | X | * | X | * | / | | | | X | | | | * | | | |
| 0 | 1.0 | * | * | * | * | * | * | * | X | * | X | X | X | * | * | / | / | * | X | X | X |
| 0 | 0.3 | | * | * | * | | | | | | | * | * | | | | | | | | |
| 0 | 0.0 | | | | | | | | | | | | | | | | | | | | |

Nota: X= identificado por ambas pruebas; *= identificado sólo por comparación modelo 1 con 2; /= identificado sólo por la comparación modelo 2 con el 3

*Tabla 3*
Ítems identificados como DIF para la estrategia 2

| | | b= -1.5 | | | | b= -1.0 | | | | b= 0.0 | | | | b= 1.0 | | | | b= 1.5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **a-dife** | **b-dife** | **a= .25** | **.60** | **.90** | **1.25** | **a= .25** | **.60** | **.90** | **1.25** | **a= .25** | **.60** | **.90** | **1.25** | **a= .25** | **.60** | **.90** | **1.25** | **a= .25** | **.60** | **.90** | **1.25** |
| 1.0 | 1.0 | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| 1.0 | 0.0 | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| .25 | 0.3 | X | | | | X | | | | X | X | | X | X | X | X | | X | X | | |
| .25 | 0.0 | X | X | X | X | X | X | X | X | X | | | | X | | | | X | | | |
| 0 | 1.0 | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | | X | X | X |
| 0 | 0.3 | | | | | | | | | | | | | | | | | | | | |
| 0 | 0.0 | | | | | | | | | | | | | | | | | | | | |

debido a que conforme el parámetro de discriminación es más elevado, el área entre las curvas asociada a la diferencia entre el grupo de referencia y el grupo focal es menor. Así, cuando la diferencia manipulada en el parámetro *a* fue de 1 generando un ítem con DIF no-uniforme simétrico (a-dife= 1 y b-dife= 0), encontramos que el área entre las CCIs del grupo de referencia y del grupo focal, calculada mediante las medidas de área de Raju (1988), fue de 2.606 cuando el parámetro de discriminación del grupo de referencia fue de 0.25, de 0.848 cuando el parámetro *a* fue de 0.6, 0.476 cuando fue de 0.9 y de 0.290 cuando $a_R$=1.25. Esta pauta también se encontró cuando las diferencias entre $a_R$ y $a_F$ fueron más pequeñas, independientemente de trabajar con ítems de mayor o menor dificultad. Cuando el DIF generado fue no-uniforme asimétrico, es decir, diferencias entre parámetros de discriminación pero también entre parámetros de dificultad, encontramos una pauta similar a la encontrada cuando el DIF manipulado fue no-uniforme simétrico; sólo cuando las diferencias en el parámetro de dificultad fueron pequeñas (b-dife=0.3), el área entre las CCIs era menor en los ítems más discriminativos. Así cuando a-dife=0.25 y b-dife=0.3, el área entre las CCIs fue de 1.648 cuando $a_R$=0.25 y 0.303 cuando $a_R$=1.25.

Si consideramos los resultados obtenidos en función del tipo de DIF manipulado, encontramos que cuando el DIF manipulado fue no-uniforme (simétrico y asimétrico) la estrategia 2 de análisis, es decir, la prueba conjunta de DIF uniforme y no-uniforme presentó en promedio un mayor porcentaje de ítems con DIF correctamente identificados (70%) frente al resto de estrategias analíticas. Así, la prueba de DIF no uniforme de la estrategia 1 y de la estrategia 3, detectaron un 33.75% de los ítems con DIF.

Cuando el DIF manipulado fue no-uniforme simétrico (diferencias sólo en el parámetro de discriminación) en promedio encontramos que tanto la estrategia 1 como la 3 identificaron correctamente un 55% de los ítems con este tipo de DIF, frente al 65% de la estrategia 2. Además, se observó que en la situación de mayor cantidad de DIF las distintas estrategias analíticas presentaron porcentajes de IC similares (ver Figura 1), tanto si la prueba estadística comprobaba la presencia de DIF uniforme como de DIF no-uniforme. Cuando la cantidad de DIF fue menor, la prueba de DIF no-uniforme de Wald-XG sobre el modelo completo (Estrategia 3-DIF no-uniforme) detectó un porcentaje de ítems con DIF (IC= 35%) menor al detectado por la prueba conjunta (Estrategia 2) de DIF (IC= 55%).

*Tabla 4*
Ítems identificados como DIF para la estrategia 3

| | | b= -1.5 | | | | b= -1.0 | | | | b= 0.0 | | | | b= 1.0 | | | | b= 1.5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **a-dife** | **b-dife** | **a= .25** | **.60** | **.90** | **1.25** | **a= .25** | **.60** | **.90** | **1.25** | **a= .25** | **.60** | **.90** | **1.25** | **a= .25** | **.60** | **.90** | **1.25** | **a= .25** | **.60** | **.90** | **1.25** |
| 1.0 | 1.0 | X | X | X | * | X | | * | * | X | * | * | X | * | / | X | X | X | / | X | X |
| 1.0 | 0.0 | X | X | X | X | X | X | X | X | X | X | X | X | * | X | X | | * | | | |
| .25 | 0.3 | X | X | | | X | | | | X | | | | * | | | | * | | | |
| .25 | 0.0 | / | / | | / | / | | / | | X | | | | * | | | | * | | | |
| 0 | 1.0 | * | | | | | | | / | | / | / | X | | | | X | | / | X | X |
| 0 | 0.3 | | | | | | | | | | | | | | | | | | | | |
| 0 | 0.0 | | | | | | | | | | | | | | | | | | | | |

Nota: X= significación estadística de ambos coeficientes; *= significación estadística sólo DIF uniforme; /= significación estadística sólo DIF no-uniforme

*Tabla 5*
Resumen del porcentaje de detección correcta del DIF

| **a-dife** | **b-dife** | **$G^2$-Unif.** | **$G^2$-NoU** | **$G^2$-C** | **Wald(G)** | **Wald(*X*G)** |
|---|---|---|---|---|---|---|
| 1.00 | 0.00 | 75% | 75% | 75% | 75% | 75% |
| 0.25 | 0.00 | 50% | 35% | 55% | 15% | 35% |
| 1.00 | 1.00 | 95% | 65% | 100% | 85% | 65% |
| 0.25 | 0.30 | 45% | 20% | 50% | 30% | 20% |
| 0.00 | 1.00 | 100% | 45% | 95% | 25% | 45% |
| 0.00 | 0.30 | 10% | 0% | 5% | 0% | 0% |
| 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |

$G^2$-Unif: Estrategia 1 de análisis DIF uniforme.
$G^2$-NoU: Estrategia 1 de análisis DIF no-uniforme.
$G^2$-C: Estrategia 2 de análisis DIF uniforme y no-uniforme conjunto.
Wald (G): Estrategia 3 de análisis DIF uniforme.
Wald(*X*G): Estrategia 3 de análisis DIF no-uniforme.

Cuando el DIF simulado fue no-uniforme asimétrico (diferencias entre el grupo focal y el de referencia, tanto en el parámetro de discriminación como en el de dificultad), fueron más eficaces tanto las pruebas de detección del DIF uniforme como la prueba conjunta de DIF uniforme y no-uniforme, con porcentajes de identificación del 100% o muy cercanos a este valor, en la condición de mayor cantidad de DIF (1.0), frente a las del DIF no-uniforme (IC= 65%). Esta pauta se mantuvo independientemente de la cantidad de DIF simulada, aunque por supuesto cuando el DIF fue menor el número de ítems detectados con DIF fue también menor.

Por último, cuando el DIF simulado fue uniforme, la prueba de DIF uniforme mediante la comparación de modelos (Estrategia 1-DIF uniforme) y la prueba conjunta de DIF uniforme y no-uniforme (Estrategia 2) mostraron las mayores tasas de IC; por contra el estadístico de Wald-G (Estrategia 3-DIF uniforme) detectó correctamente un menor número de ítems con DIF. Así, en promedio un 55% de los ítems fueron detectados correctamente por la prueba de

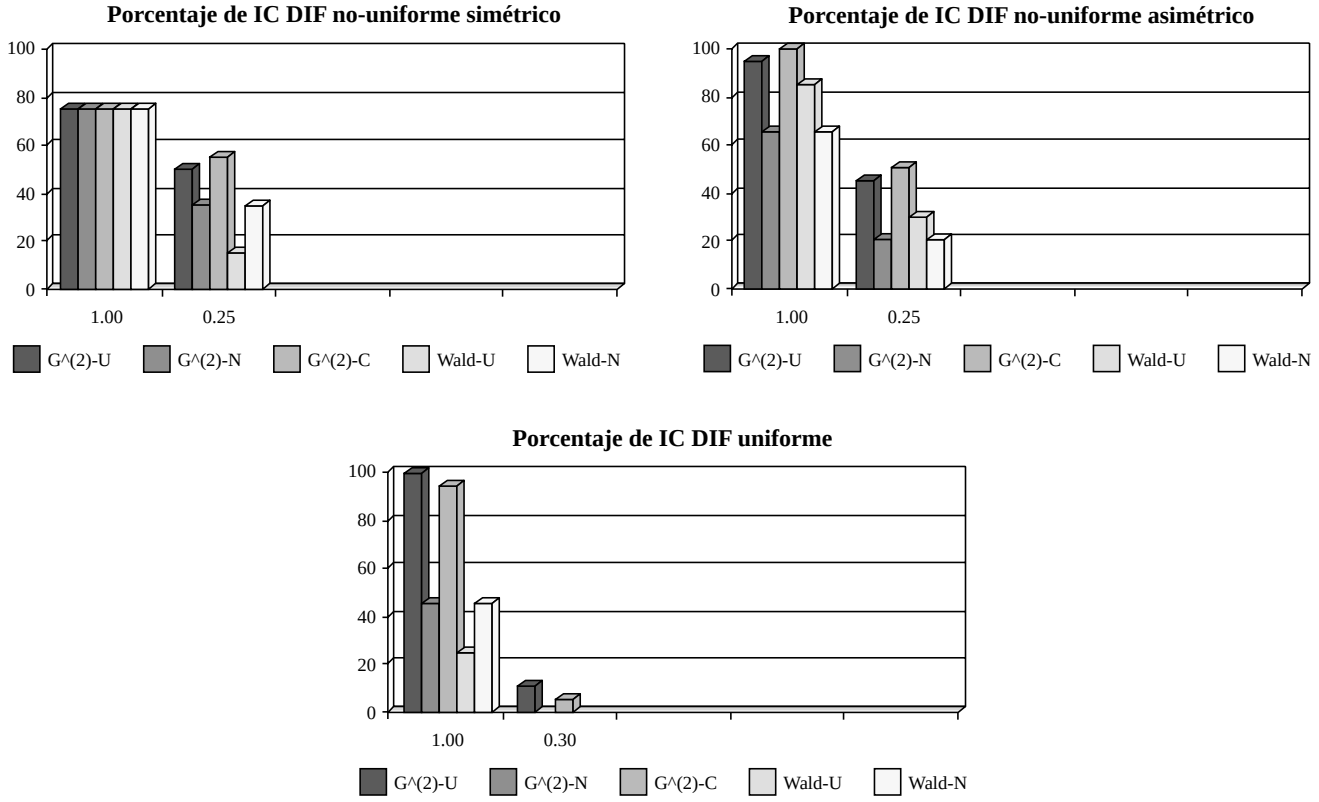


***Figura 1.*** *Porcentaje de IC en las distintas estrategias de análisis de la regresión logística*

DIF uniforme de la estrategia 1, un 50% por la prueba conjunta de DIF (estrategia 2) y un 12.5% por la prueba de DIF uniforme de la estrategia 3. Por otro lado, tal y como era de esperar, las distintas estrategias analíticas presentaron tasas de IC más altas cuando la cantidad de DIF simulado fue mayor, siendo del 100% cuando se utilizó la prueba de DIF uniforme de la estrategia 1, y del 95% cuando se evaluó el DIF con la prueba conjunta de DIF (estrategia 2), estas tasas de IC fueron muy bajas, hasta del 0%, cuando la cantidad de DIF uniforme fue menor (ver Tabla 5).

En cuanto a los ítems de no-DIF los tres procedimientos controlaron las identificaciones incorrectas, siendo este porcentaje de cero.

## Discusión y conclusiones

Los resultados encontrados en este estudio apuntan a que, cuando el DIF simulado fue no-uniforme simétrico, las tres estrategias de análisis presentaron un nivel de eficacia similar, siendo menos costoso computacionalmente ajustar el modelo completo y comprobar la significación o no de la interacción habilidad x grupo mediante el estadístico de Wald. Cuando el DIF simulado fue no-uniforme asimétrico, la prueba conjunta de DIF uniforme y no-uniforme aparece como la estrategia más eficaz, ya que presentó las tasas de IC más altas, frente a la prueba individual de DIF no-uniforme. La prueba de DIF uniforme también resultó efectiva en la identificación de este tipo de DIF.

Por otro lado, cuando el DIF simulado fue uniforme, probar el efecto de la variable de agrupamiento mediante la comparación del modelo que la incluye con el modelo que implica sólo a la habilidad de los sujetos, resultó ser la estrategia más efectiva, seguida en grado de eficacia por la prueba conjunta de DIF uniforme y no-uniforme.

Tal y como era de esperar, las distintas estrategias analíticas resultaron más efectivas en la detección del DIF cuando la cantidad de DIF fue mayor, en línea con los estudios previos (Hidalgo y Gómez, 2000; Narayanan y Swaminathan, 1996; Rogers y Swaminathan, 1993) sobre el efecto de esta variable en la capacidad de detección del DIF con regresión logística.

Por último, ninguno de los procedimientos utilizados detectó DIF en los ítems simulados sin DIF, presentando un buen control de la tasa de error tipo I.

En términos generales, si atendemos tanto al criterio de mayor potencia en la detección del DIF como de menor costo computacional, se recomienda a los psicólogos y educadores prácticos que analizan, traducen y adaptan tests, el uso del análisis de regresión logística implementando la estrategia 2, que somete a comprobación la presencia conjunta de DIF uniforme y no-uniforme. Esta estrategia de análisis, resultó ser relativamente la más eficaz en la identificación de ítems con DIF, requiriendo el ajuste de dos modelos, el de ausencia de DIF y el modelo completo. Sin embargo, la limitación más importante de esta alternativa de análisis, es que no permite conocer el tipo de DIF que presenta el ítem, y requiere de un estudio posterior del mismo. En estos casos la representación gráfica de las probabilidades de respuesta al ítem para los grupos bajo estudio suele ser una alternativa que permite diagnosticar el tipo de DIF para un ítem determinado. Otra posibilidad es comprobar en el modelo completo, mediante el estadístico de Wald, la significación estadística de la interacción (estrategia 3-DIF no-uniforme): si este tér-

mino resultase estadísticamente significativo se podría presuponer la existencia de DIF no-uniforme.

En cualquier caso, es necesario más investigación para ampliar la capacidad de generalización de los resultados encontrados aquí, por ejemplo, analizando la eficiencia de estas estrategias en otras condiciones tales como un mayor porcentaje de ítems con DIF en el test, así como conocer el correspondiente efecto de emplear procedimientos de purificación del test (Hidalgo y Gómez, 2003) en cada una de las estrategias de análisis.

## Notas

[1] El valor de la verosimilitud es -2 veces el logaritmo de la verosimilitud, y se suele representar por -2LL o por D (desvianza).

[2] $G^2$= D (modelo con la variable) – D (modelo sin la variable)= -2 log (verosimilitud modelo con la variable / verosimilitud modelo sin la variable).

[3] El estadístico de Wald es igual al cuadrado del cociente entre el coeficiente y su error estándar.

## Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Ciencia y Tecnología y los Fondos FEDER (Proyecto n.º: BSO2001-3751-C02-02).

Parte de este trabajo fue presentado al VII Congreso de Metodología de las Ciencias del Comportamiento, Madrid, septiembre. del 2001.

## Referencias

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. En P.W. Holland y H. Wainer (eds.): *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum.

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

Clauser, B.E. y Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-44.

Clauser, B.E., Nungester, R.J., Mazor, K. y Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33*, 202-214.

De Ayala, R.J., Kim, S.H., Stapleton, L.M. y Dayton, C. (1999). *A reconceptualization of Differential Item Functioning*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canadá.

Ferreres, D., Fidalgo, A. y Muñiz, J. (2000). Detección del funcionamiento diferencial de los ítems no uniforme: comparación de los métodos Mantel-Haenzel y regresión logística. *Psicothema, 12*, 220-225.

Gierl, M.J., Gotzmann, A. y Boughton, K.A. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education, 17*, 241-264.

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

Hidalgo, M.D. y Gómez, J. (2000). Comparación de la eficacia de la regresión logística politómica y análisis discriminante logístico en la detección del DIF no uniforme. *Psicothema, 12*, 298-300.

Hidalgo, M.D. y Gómez, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment, 19*, 1-11.

Hidalgo, M.D. y López, J.A. (2004). DIF detection and effect size: a comparision between logistic regression and Mantel-Haenszel variation. *Educational and Psychological Measurement, 64*, 903-915.

Hosmer, D.W. y Lemeshow, S. (1989). *Applied Logistic Regression*. New York, NY: Wiley.

Mellenbergh, G.J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105-108.

Miller, M.D. y Linn, R.L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement, 25*, 205-219.

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

Narayanan, P. y Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257-274.

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

Rogers, H.J. y Swaminathan, H. (1993). A comparison of Logistic Regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116.

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