Psicothema, 1996. Vol. 8, nº 1, pp. 173-180
ISSN 0214 - 9915 CODEN PSOTEG

SOFTWARE, INSTRUMENTACIÓN Y METODOLOGÍA

DETECCIÓN DE FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS MEDIANTE ANÁLISIS DE RESIDUALES: UNA APLICACIÓN DE LA TRI

Pedro Prieto Marañón* y Mª Isabel Barbero García
* Universidad de La Laguna, ** U.N.E.D.

En el presente trabajo se expone un procedimiento de detección del funcionamiento diferencial de los ítems (DIF) mediante un programa de análisis de residuales elaborado a tal efecto (RES.BAS) que, dentro del marco de la TRI, y siguiendo la propuesta de Linn y Harnisch (1981), permite comprobar el ajuste entre los datos del grupo focal y un modelo previamente determinado en el grupo de referencia. Para estimar lo adecuado de tal procedimiento y la utilidad del programa se ha comparado, mediante un estudio de simulación, con otros cuatro métodos de detección de DIF: Mantel-Haenszel, prueba χ^2 de Lord, modelos log-lineales y regresión logística. Los resultados muestran que tanto el método propuesto, como la regresión logística, resultan ser más eficaces que los de Mantel-Haenszel y Lord y ligeramente mejores con respecto al modelo log-lineal. Una de las principales ventajas del procedimiento expuesto es su capacidad para diferenciar no sólo entre DIF uniforme y no uniforme, sino para mostrar la dirección, intensidad y niveles del rasgo en los que aparece el efecto del DIF.

Detecting differential item functioning through residual analysis: an irt applicatton. In this paper we present an IRT procedure for DIF detection through a program of residual analysis (RES.BAS) based on the Linn & Harnsich (1981) method of analizing residuals obtained when fitting data from the focal group to a previously determined model on a reference group. For estimating accuracy of this procedure, a comparison with four other methods - Mantel-Haenszel, Lord`s test, log-lineal model and logistic regression- has been carried through a simulation study. Results show that both the proposed method and the logistic regression one, are superior to Mantel-Haenszel and slightly better than the log-lineal model. One of the main advantages of residual program is its capacity, not only to differentiate between uniform and no uniform DIF, but also in detecting direction, size and ability levels in which DIF is present.

La detección de ítems sesgados, o que presentan un funcionamiento diferencial (DIF), ha sido objeto de gran atención en los últimos años, convirtiéndose en una importante área dentro del campo de la aplicación de pruebas psicométricas. Los estudios acerca del DIF pretenden comparar la ejecución relativa del subgrupo de interés, o grupo focal, frente a un grupo de

Correspondencia: Pedro Prieto Marañón
Facultad de Psicología
Universidad de La Laguna
Tenerife. Spain

referencia. Podemos hablar de la presencia de DIF cuando miembros de ambos grupos con igual nivel de competencia o habilidad poseen diferentes probabilidades de acertar o superar el ítem en cuestión (Lord, 1980). Aunque diversos autores, como es el caso del propio Lord, han empleado el término *sesgo* para referirse al DIF, esta terminología resulta algo confusa (Donoughe y Allen, 1993). El término *sesgo* implica el que un grupo posee una "injusta" ventaja sobre el otro. Sin embargo, los grupos pueden presentar diferencias entre ellos en sus respuestas ante un ítem debido a razones diferentes al sesgo.

Las técnicas estadísticas son apropiadas para detectar si un ítem funciona de manera diferente en dos grupos; pero en ningún caso ofrecen información acerca de si esta diferencia es producto legítimo del constructo objeto de medición. Si el DIF observado puede calificarse o no como sesgo es ya una cuestión de validez de constructo.

Para la detección del DIF se han propuesto numerosos métodos que abarcan desde aproximaciones centradas en el análisis de varianza, hasta el empleo de las técnicas derivadas de la teoría de la respuesta al ítem (Shepard y Camilli, 1994). En la actualidad, algunos de los más prometedores métodos (Swaminathan y Rogers, 1990; Hambleton, Swaminathan y Rogers, 1991) parecen ser aquellos basados en los principios de la TRI junto con el método de Mantel-Haenszel propuesto por Holland y Thayer (1988).

Este último procedimiento resulta particularmente atractivo debido a su facilidad de implementación y al hecho de poseer un test de significación asociado. Sin embargo, a pesar de lo práctico que pueda resultar, el procedimiento de Mantel-Haenszel no resulta útil ni adecuado en la detección de DIF no uniforme (Hambleton y Rogers, 1989, Rogers y Swaminathan, 1990).

Los términos DIF uniforme y no uniforme fueron acuñados por Mellenbergh (1982), considerando que existe DIF uniforme cuando no existe interacción entre el nivel de aptitud y la pertenencia a un grupo, es decir, cuando la probabilidad de contestar correctamente a determinado ítem es uniformemente superior para uno de los grupos a lo largo de los diversos niveles de aptitud. Por su parte, existe DIF no uniforme cuando sí existe interacción entre nivel de aptitud y grupo, es decir, la diferencia en la probabilidad de la respuesta correcta no es la misma a lo largo de todos los niveles de aptitud. (Figs. 1a y b).

Rogers y Swaminathan (1993) distinguen a su vez entre DIF no uniforme y no uniforme mixto. Este último termino hace referencia a aquellos casos en los que $a_i \neq a_j$ y $b_i \neq b_j$, frente a aquellos en los que únicamente $a_i \neq a_j$. (Fig. 1c).

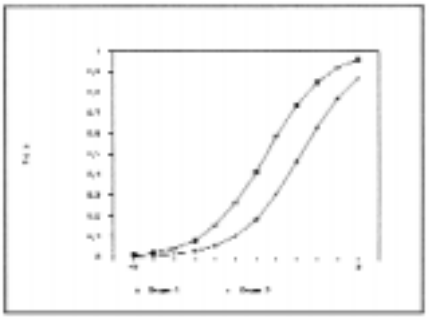

Figura 1a. DIF uniforme

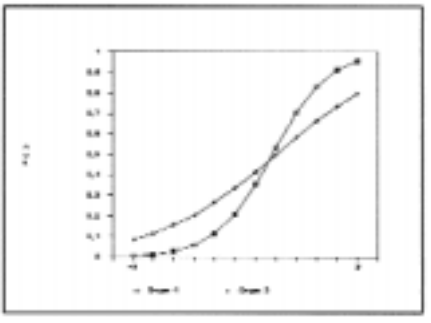

Figura 1b. DIF no uniforme

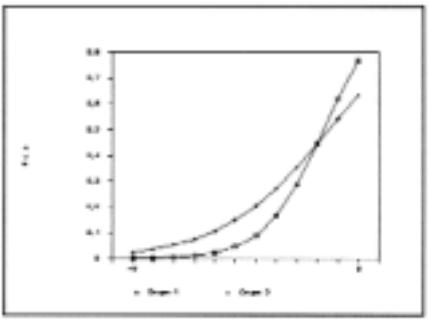

Figura 1c. DIF no uniforme mixto

Diversos investigadores (Camilli, 1979; Marascuillo y Slaughter, 1981; Mellenbergh, 1982) han sugerido el uso de modelos log-lineales como un medio eficiente de detectar la presencia de DIF no uniforme. Según este modelo, las respuestas a cada ítem pueden ser predichas a partir de la pertenencia a uno de los grupos, el nivel de aptitud y la interacción de ambos factores. Este es un procedimiento análogo a un análisis de varianza en el cual los niveles de aptitud son considerados como categorías discretas no ordenadas. Si el modelo que mejor ajusta a los datos en ambos grupos incluye únicamente los efectos de nivel de aptitud y del término constante, entonces se concluye que no existe DIF. Si además es necesario incluir el efecto del grupo estamos ante un caso de DIF uniforme. Si, por último, se debe incluir el efecto de la interacción nivel x grupo, podemos hablar de DIF no uniforme.

Una de las principales críticas a esta técnica la presentan Swaminathan y Rogers (1990) quienes consideran que el ignorar la naturaleza ordinal de los niveles de aptitud supone desperdiciar buena parte de la información disponible. En su opinión, un método que considere la aptitud como una variable continua puede resultar más adecuado.

Como alternativa a estos métodos, y también frente a los más costosos métodos derivados de la TRI, Swaminathan y Rogers (1990) proponen un procedimiento de regresión logística para el análisis del DIF que resulta efectivo en la detección tanto de DIF uniforme como no uniforme.

El modelo de regresión logístico para la detección de DIF propuesto por estos autores, viene expresado por:

$$P(u = 1) = \frac{e^{Z}}{(1+e^{Z})}$$

dónde

$$Z = \pi_0 + \pi_1\theta + \pi_2 g + \pi_3 (\theta g)$$

En este modelo θ es el nivel de aptitud observado de un individuo (por lo general, la puntuación total en el test), y *g* representa la pertenencia a uno de los grupos, con un valor de *1* para el grupo de referencia y de *0* para el grupo focal. Por su parte θg es el producto de las dos variables independientes θ y *g*.

A partir de esta codificación, el parámetro π_2 corresponde a la diferencia entre los grupos en la ejecución del ítem, y el parámetro π_3 corresponde a la interacción entre aptitud y grupo. Un ítem mostrará DIF uniforme si $\pi_2 \neq 0$ y $\pi_3 = 0$, y DIF no uniforme si $\pi_3 \neq 0$, independientemente del valor de π_2. Swaminathan y Rogers (1990), así mismo, proponen un estadístico para comprobar la hipótesis nula de la simultaneidad de $\pi_2 = 0$ y $\pi_3 = 0$. Este estadístico se ajusta a una distribución χ^2 con 2 grados de libertad.

Uno de los inconvenientes que ofrecen estos métodos estriba en el cuestionable uso de la puntuación total en el test como una medida del nivel de aptitud, dado que esta puntuación no es perfectamente fiable. Es, por contra, la naturaleza de invarianza de los parámetros y del valor de θ con respecto de las muestras utilizadas una de las razones que hace que los modelos

de la TRI se presenten como procedimientos con una mayor solidez teórica y complejidad estadística en la detección de DIF (Petersen, 1977; Lord, 1980).

Entre los diferentes procedimientos enmarcados dentro de la TRI, Linn y Harnisch (1981) propusieron un índice basado en los residuales X_j - $P(\theta_j)$, dónde X_j es la respuesta observada del individuo *j*. El procedimiento (Shepard, Camilli y Williams, 1985) consiste en determinar, en primer lugar, un modelo que se ajuste adecuadamente a los datos del primer grupo (o combinación de ambos). Posteriormente se estima un índice global de los residuos estandarizados en el grupo focal con respecto a la curva característica del de referencia, mediante la expresión:

$$B = \sum_{j_1}^{N_2} \frac{X_j - P(\theta_j)}{\sqrt{P(\theta_j)(1-P(\theta_j)}}$$

Un valor de *B* próximo a cero indica que el modelo escogido también ajusta con los datos del segundo grupo. Un valor absoluto diferente de cero refleja la presencia de DIF.

Un procedimiento basado en este modelo ha sido empleado por Barbero (1994) en un estudio acerca de la posible existencia de DIF en una prueba de rendimiento en Ciencias entre diversas Comunidades Autónomas.

El procedimiento seguido consistió, en primer lugar, en determinar un modelo de TRI que ajustase los datos de los dos grupos a comparar. Una vez elegido un modelo con un buen índice de ajuste, se estimaron los valores de θ y de los parámetros -estandarizados- en ambos grupos. Posteriormente, mediante el módulo de análisis de residuos (programa RES.BAS) del programa MABEL (Prieto et al, 1994) se estimó el ajuste entre los datos -respuestas empíricas- del segundo grupo y el modelo definido por los parámetros -estandarizados- del primer grupo.

A través de los índices de ajuste que ofrece este programa - proporción de residuos absolutos estandarizados menores o igual que 1.96 (α = 0.05), residuo medio, residuo absoluto medio y χ^2 (Wright y Panchapakesan, 1969)- se consideró que presentaban DIF aquellos ítems que no poseían un adecuado nivel de ajuste. Una de las ventajas que aporta este procedimiento es que, dado que la salida del programa ofrece una gráfica de la distribución de los residuales a lo largo de las categorías en que se divide θ para su cómputo, dicha gráfica -junto con el valor del residuo estimado en cada categoría- nos permite determinar, no sólo si se trata de DIF uniforme o no uniforme, sino incluso la dirección, valor y niveles en los que se presentan las diferencias.

En las figuras 2a, b, c y d se presentan varios ejemplos de este procedimiento. Como puede observarse la figura 2a, representa un item en el que no se presenta DIF dado que el ajuste es bueno (χ^2 = 3.35; Prop. RSD = 1) a lo largo de las 12 categorías en las que se dividió θ. En cambio, en la figura 2b se observa cómo el ítem no se ajusta al modelo (χ^2 = 83.87; Prop. RSD = .50). La distribución uniforme de los residuos mayores que 1.96, nos indica que se trata de un caso de DIF uniforme, siendo en este caso más dificil para los miembros del grupo focal que para los del de referencia, dado que la diferencia entre las proporciones teórica y empírica de aciertos es positiva a lo largo de las categorías. La gráfica 2c nos ofrece otro caso de DIF (χ^2 = 145.70; Prop. RSD = .42), sin embargo, en esta ocasión el DIF no es uniforme, siendo el ítem en cuestión más fácil para los sujetos del grupo focal con un nivel de aptitud medio-bajo y más difícil para los de nivel alto. Por último, en la figura 2d se presenta una distribución de

residuales para el caso de DIF no uniforme mixto, la cual se asemeja bastante al caso de DIF uniforme.

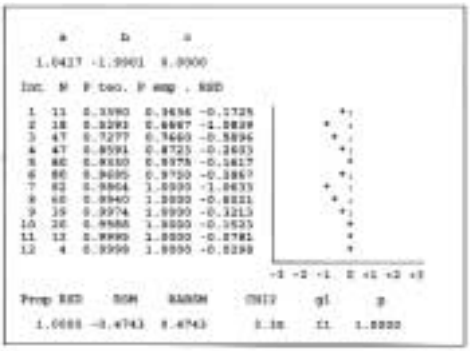

Figura 2a. Distribución de residuales. Ausencia de DIF

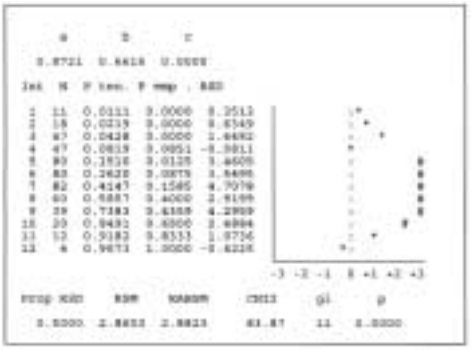

Figura 2b. Distribución de residuales DIF uniforme

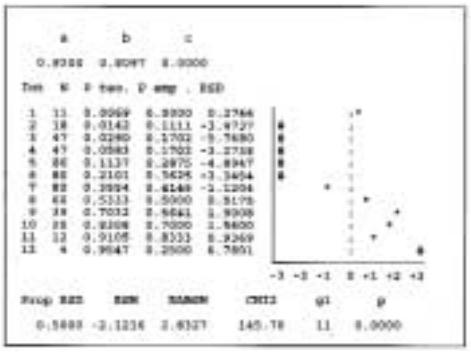

Figura 2c. Distribución de residuales DIF no uniforme

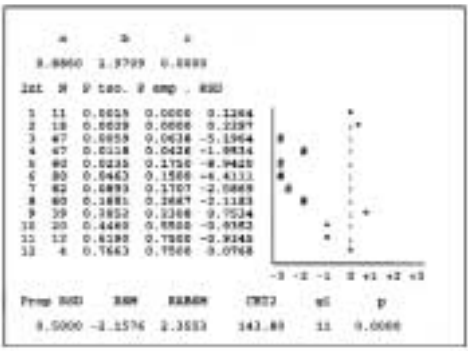

Figura 2d. Distribución de residuales. DIF no uniforme mixto

Con objeto de comprobar la utilidad del procedimiento utilizado por Barbero (1994) presentamos a continuación un estudio en el que es comparado con otros procedimientos.

Método

Mediante el programa SIMTRI (San Luis et al, 1994) se generó, a partir de la función logística de 2 parámetros, una matriz de respuestas simuladas para 50 ítems y 500 sujetos. Los valores de θ se generaron a partir de una distribución normal N(0,1). Por su parte los valores de *a* y *b* se generaron a partir de distribuciones uniformes entre 1 y 2 y -2 y +2 respectivamente. Con estos mismos valores se generó una nueva matriz de aciertos/errores para una segunda muestra, variando únicamente los valores de los parámetros *a* y *b* para los 6 primeros ítems. Los nuevos valores en este segundo grupo fueron:

$a_{12} = a_{11}$	$b_{12} = b_{11} + .64$ (DIF uniforme).
$a_{22} = a_{21}$	$b_{22} = b_{21} - .48$ (DIF uniforme).
$a_{32} = a_{31} - .5$	$b_{32} = b_{31}$ (DIF no uniforme).
$a_{42} = a_{41} + 1$	$b_{42} = b_{41}$ (DIF no uniforme).
$a_{52} = a_{51} - .5$	$b_{52} = b_{51} + .64$ (DIF no uniforme mixto).
$a_{62} = a_{61} + 1$	$b_{62} = b_{61} - .48$ (DIF no uniforme mixto).

Los valores de las diferencias de *a* y *b* se tomaron, siguiendo a Swaminathan y Rogers (1990), de manera que reflejasen valores de DIF entre moderados y altos.

Los datos así generados fueron posteriormente sometidos a cinco diferentes métodos de detección de DIF: χ^2 de Mantel-Haenszel, prueba χ^2 de Lord (Lord, 1980), modelos log-lineales, regresión logística y el método de residuales mediante el programa RES.BAS. En cada caso se tomó como nivel de significación estadística un valor de $\alpha = .05$.

Resultados

En la Tabla 1 presentamos las tasas de detección de los cinco métodos. Indiquemos, antes que nada, que para el análisis mediante los métodos basados en la TRI (prueba de Lord y residuales), se ajustó el modelo logístico de dos parámetros en ambas muestras, siendo el nivel de ajuste del modelo, medido a través de la proporción de residuos estandarizados absolutos menor o igual que 2, de .9817 en el primer grupo y .9800 en el segundo.

Tabla 1
Comparación de los cinco métodos en la detección de DIF

Item	Nº de ítems detectados como sesgados					
	Nº	MH	Lord	Log-L	R. Log	Res
DIF	6	4	6	6	5	6
Falsos posit.	42	0	8	2	2	3

Como se aprecia en los resultados, la tasa de ítems con DIF correctamente detectados mediante el método de Mantel-Haenszel es ligeramente inferior al resto, siendo ésta muy similar en los restantes procedimientos. Por contra, la tasa de falsos positivos es nula para este método, al tiempo que es elevada cuando se aplica el estadístico χ^2 de Lord, mientras que en el resto de los procedimientos se mantiene entre los 2 y 3 que cabría esperar por puro azar para un nivel de $\alpha = .05$.

En cuanto a estos falsos positivos, todos los detectados por los métodos capaces de diferenciar entre DIF uniforme y no uniforme, señalan a estos como ítems con DIF no uniforme. Tal vez sea esta la razón de la nula tasa de falsos positivos de Mantel-Haenszel, producto de su baja capacidad, como señalan Hambleton y Rogers (1989) para detectar este tipo de DIF.

Como corroboración de esta baja capacidad para la detección de DIF no uniforme, se encuentra el hecho de que únicamente fueron detectados por Mantel-Haenszel los dos ítems con DIF uniforme y los dos con DIF no uniforme mixto, dada la fácil confusión entre ambos tipos de DIF, tal como apuntan Rogers y Swaminathan (1993).

En cuanto a los tres métodos capaces de clasificar el tipo de DIF, presentamos en la Tabla 2 el número de ítems correctamente detectados en función de aquél.

Tabla 2
Comparación de los tres métodos que discriminan entre DIF uniforme y no uniforme

DIF	Nº de ítems detectados como sesgados			
	Nº	Log-L	R. Log	Res
Uniforme	2	0	1	2
No uniforme	2	2	2	2
No uniforme mixto	2	1	2	1

En los ítems con DIF uniforme, el único procedimiento que los identifica correctamente es el basado en el análisis de residuales mediante el programa RES.BAS dándose el caso de que el modelo log-lineal lo hace de forma totalmente incorrecta.

En el caso de DIF no uniforme, este es detectado correctamente por los tres pro-

cedimientos, cuando este es estrictamente no uniforme. Cuando se trata de DIF no uniforme mixto, el modelo de regresión logística es el único en detectar correctamente los dos casos, mientras que los otros dos procedimientos catalogan el DIF del ítem 5 como uniforme.

Conclusiones

En el presente trabajo se expone un procedimiento de detección de DIF mediante un programa elaborado a tal efecto (programa RES.BAS), que dentro del marco de la TRI, y siguiendo la propuesta de Linn y Harnisch (1981), se basa en el análisis de los residuales resultantes de comprobar el ajuste entre los datos del grupo de interés y un modelo previamente determinado en el grupo de referencia.

Para estimar lo adecuado de tal procedimiento se ha comparado, en un estudio de simulación, con otros cuatro métodos de detección de DIF: Mantel-Haenszel, prueba χ^2 de Lord (Lord, 1980), modelos log-lineales y regresión logística. Los resultados muestran que tanto el método propuesto, como la regresión logística, resultan ser más eficaces que los de Mantel-Haenszel y Lord y ligeramente mejores con respecto al modelo log-lineal.

El procedimiento de Mantel-Haenszel, además de su incapacidad para distinguir entre DIF uniforme y no uniforme, es insensible a la presencia de este último, confirmando lo apuntado por autores como Rogers y Swaminathan (1993). Por su parte, el estadístico de Lord presenta una alta tasa de falsos positivos.

El mayor coste computacional del procedimiento basado en los residuales, comparado con el de Mantel-Haenszel y los modelos log-lineales, se ve recompensado por su capacidad para, no sólo diferenciar entre DIF uniforme y no uniforme -algo que el método de la regresión logística hace bastante bien- , sino para, incluso, mostrar la dirección, intensidad y niveles del rasgo en los que aparece el efecto del DIF, lo cual se convierte en su principal ventaja.

A la espera de futuros trabajos en los que se analice su validez y precisión bajo diversas condiciones (diferentes tamaños muestrales y de DIF, nivel de ajuste de los modelos...), los resultados encontrados en este estudio exploratorio nos hacen considerar esta aproximación como un procedimiento válido y eficiente, a través del cúal es posible, no sólo identificar aquellos ítems que presentan DIF, sino incluso conseguir una mayor comprensión de la naturaleza del mismo.

Referencias

Barbero, M.I. (1994): *Teoría de la respuesta al ítem y evaluación del rendimiento en ciencias de los niños y niñas españoles de 13 años. Desarrollo del programa GENESTE*. Trabajo de investigación presentado como ejercicio de oposición. Trabajo inédito.

Camilli, G. y Shepard, L.A. (1994): "*Methods for identifying biased test items*", London. Sage.

Camilli, G. (1979): "A critique of the chi-square methods of assesing item bias", *Laboratory of educational research*, University of Colorado, Boulder.

Donoughe, J.R. y Allen, N.L. (1993): "Thin versus thick matching in the Mantel-Haenszel Procedure for detecting DIF", *J. of Educational Statistics*, 18,2, 131-154.

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

Hambleton, R.K., Swaminathan, H. y Rogers, H.J. (1991): *Fundamentals of item response theory*, London, Sage.

Holland, P.W. y Thayer, D.T. (1988): "Differential item perfomance amd the Mantel-Haenszel procedure", en Wainer, H. y Braun, H.I. (Eds.): *Test validity* (pp 129-145). Hillsdale, Nueva Jersey.

Linn, R.L. y Harsnich, D.L. (1981): "Interacions between item content and group membreship on achievement test items", *J. of Educational Measurement*, 18, 109-118.

Lord, F.M. (1980): *Applications of item response theory to practical testing problems*, Hillsdale, Nueva Jersey, ETS.

Marascuillo, L.A. y Slaugther, R.E. (1981): "Statistical procedures for identifying sources of item bias based on chi-square statistics", *J. of Educational Measurement*, 18, 229-248.

Mellenbergh, G.J. (1982): "Contingency tables models for assesing item bias", *J. of Educational Statistics*, 7, 105-108.

Petersen, N.S. (1980): "Bias in the selection rule-bias in the test", en J.T. van der Kamp et al (Eds.) "*Psychometrics for educational debates*" (pp 103-107), Nueva York, Wiley.

Prieto, P.; Barbero, M.I.; San Luis, C. y Sánchez, J.A. (1994): "MABEL: un programa de control de BILOG y análisis de residuales". Enviado para su publicación.

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, 2, 105-115.

San Luis, C., Prieto, P., Barbero, M. y Sánchez, J.A. (1994): "SIMTRI: Un simulador para TRI". Enviado para su publicación.

Shepard, G.; Camilli, L.A. y Williams, D. "Validity for aproximation techniques for detecting item bias", *J. of Educational Measurement*, 22, 77-105.

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

Wright, B.D. y Panchapakesan, N. (1969): "A procedure for samplefree item analisys", *Educational and Psychological Measurement*, 29, 23-48.