Psicothema, 1999. Vol. 11, nº 3, pp. 691-697
ISSN 0214 - 9915 CODEN PSOTEG

DETECCIÓN DEL FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS EN UNA PRUEBA DE CIENCIAS

Pedro Prieto Marañón, M. Isabel Barbero García*
y Concepción San Luis Costas
Universidad de La Laguna y * Universidad Nacional de Educación a Distancia

Empleando los recursos del National Asessment of Educational Progress (NAEP) se ha realizado un estudio con el fin de determinar la posible presencia de funcionamiento diferencial de los ítems que componen una prueba de ciencias entre niños y niñas españoles de 13 años, como paso previo para la identificación de ítems sesgados. Para ello se emplearon una variación del procedimiento estándar de Mantel-Haenszel propuesta por Mazor, Clauser and Hambleton (1994) y un procedimiento de análisis de residuales basado en los modelos de la TRI, propuesto por Prieto, Barbero y San Luis (1996), aplicados ambos de forma iterativa. Los resultados de este trabajo sugieren que los procedimientos estadísticos para la detección del FDI pueden verse mejorados con su aplicación de forma iterativa, y que la aplicación conjunta de diferentes procedimientos puede llevar a conclusiones más fiables acerca de la identificación de posibles ítems sesgados.

Detection of differential item functioning in a science test. Using National Asessment of Educational Progress (NAEP) resources, a study was carried out to determine the presence of differential functioninig of the items of a Science Test between 13-years-old Spanish boys and girls as a first step in the identification of biased items. A variation of the standard Mantel-Haenszel procedure proposed by Mazor, Clauser and Hambleton (1994) and a residual analysis procedure based on IRT techniques, proposed by Prieto, Barbero and San Luis (1996), are used in an iterative way to detect items showing differential functioning. Results from this work suggest that statistical procedures for detecting DIF may be improved by applying them in an iterative way, and that the application of several strategies to the same pool of data may lead to obtain more reliable conclusions about the identification of possible biased items.

El estudio del sesgo en las pruebas psicométricas es una cuestión que surge paralelamente en el tiempo a la aparición del movimiento de los derechos civiles (igualdad de derechos y oportunidades) iniciado en EE.UU. a comienzos de la década de los setenta. Desde esta concepción de igualdad de oportunidades se exigen condiciones a los instrumentos de medida de variables psicológicas que garanticen que no favorecerán a los miembros de determinado grupo. Es esta condición la que hace referencia al concepto de sesgo.

Correspondencia: Pedro Prieto Marañón
Facultad de Psicología
Universidad de La Laguna
38205 La Laguna - Tenerife (Spain)
E-mail: pprieto@ull.es

El término sesgo, presenta ciertas connotaciones relacionadas con el concepto de injusticia, que han enturbiado su uso en el marco de la medición en Psicología. A fin de desligar ambos significados, sesgo en sentido psicométrico y sesgo como indicador de injusticia, diversos autores (Jensen, 1980; Angoff, 1982; Holland y Thayer, 1988; Fidalgo, 1996, 1996b) desarrollaron una serie de trabajos que han dado lugar a la definición de nuevos términos que hoy día nos resultan familiares: *FDI (*Funcionamiento Diferencial de los Items) e *Impacto* cuyo objetivo es el disponer de un significado exclusivamente métrico y alejado de cualquier otra connotación. El término FDI fue acuñado por Holland y Thayer (1988) y describe a los ítems que tienen propiedades estadísticas distintas en diferentes grupos, propiedades que dan lugar a diferencias que en realidad no se corresponden con comportamientos diferentes de la variable medida. Así se entiende que un ítem presenta FDI cuando la probabilidad de resolverlo correctamente es diferente para sujetos que perteneciendo a distintos grupos tienen el mismo nivel de habilidad. Por el contario, el término *impacto*, según la definición de Ackerman (1992), consiste en una diferencia que se presenta entre grupos en la resolución de un ítem y que se debe a variaciones reales en la variable medida. Así, se dice que existe *impacto*, cuando los sujetos que tienen el mismo nivel de habilidad, con independencia del grupo al que pertenezcan, tienen la misma probabilidad de responder correctamente al ítem y, sin embargo, hay diferencias en la actuación media de los grupos en el ítem. El término *sesgo*, por su parte, se emplea hoy en día cuando se hacen juicios de valor acerca de las causas y/o explicaciones del comportamiento diferencial de los ítems (FDI) detectado entre distintos grupos. Diferentes autores han propuesto distintas explicaciones a la problemática del sesgo, así Ackerman, (1988, 1993) o Shealy y Stout, (1993), entre otros, señalan que puede deberse a un problema de validez de constructo de la prueba. Otros autores, entre los que se puede citar a Cole y Moss (1989), sitúan el problema del sesgo dentro del marco de la validez diferencial. La *teoría de los violadores potenciales* propuesta por Oort (1992, 1993) proporciona un marco que permite considerar de forma simultánea el conjunto de factores que pueden influir en la actuación de los sujetos en el ítem y pueden ayudar a comprender porqué en ocasiones sujetos que tienen el mismo nivel de habilidad poseen distinta probabilidad de responder correctamente a un ítem que supuestamente mide esa habilidad. Estos violadores potenciales o rasgos sesgadores se pueden definir en términos de los contenidos y procesos mentales implicados en la tarea implícita en el ítem y se relacionan, fundamentalmente, con el nivel educativo, social y económico de los sujetos pertenecientes a distintos grupos (Muñiz, 1990).

Teniendo esta idea en mente, y como un paso más en el proceso de baremación de una prueba de Ciencias adaptada del ámbito anglosajón, hemos llevado a cabo el presente estudio con el fin de determinar la presencia de ítems en la prueba que funcionen de forma diferencial (FDI) en los grupos comparados. El fin último que se persigue es el establecer si las diferencias encontradas en rendimiento en dicha prueba entre niños y niñas de 13 años de algunas de las Comunidades Autónomas españolas se deben a diferencias reales (impacto) o, por el contrario, a algún tipo de sesgo.

Procedimiento

Muestra

La muestra utilizada en este trabajo está compuesta de 1.756 sujetos de los cuales 868 eran niños y 888 niñas. Se trata de la muestra española utilizada en el primer es-

tudio internacional (*International Assessment of Educational Progress)* llevado a cabo por el ETS dentro del programa NAEP.

Prueba utilizada

Los 60 ítems que componían la prueba de Ciencias empleada en este estudio fueron seleccionados de un banco de 188 ítems utilizados en 1986 en el programa de evaluación del sistema educativo americano (NAEP) y estaban repartidos en las siguientes categorías en función de sus contenidos: 15 ítems de Ciencias Naturales, 10 de física, 9 de Química, 8 de Ciencias de la Tierra y del Espacio y 13 de Fundamentos de la Ciencia.

Análisis Estadísticos Previos.

En primer lugar se llevaron a cabo una serie de análisis estadísticos con el fin de averiguar la posible existencia de ítems en los que existieran diferencias significativas en las respuestas de los sujetos de los dos grupos. Al hacer las comparaciones entre los niños y niñas se encontraron diferencias significativas ($p < .01$) en 13 de los 60 ítems; en 10 de ellos la diferencia era favorable a los niños y en 3 a las niñas. Antes de llegar a ninguna conclusión acerca del superior rendimiento de los niños, se consideró conveniente realizar un estudio de FDI que pudiera poner de manifiesto la posible existencia de algún sesgo en el test.

Estudio del FDI

De entre los diversos procedimientos estadísticos desarrollados para la detección del FDI se utilizó una variación del procedimiento de Mantel-Haenszel (MH) propuesta por Mazor, Clauser y Hambleton (1994) y un procedimiento de análisis de residuales (Linn y Harnisch, 1981) basado en las técnicas de la TRI propuesto por Prieto, Barbero y San Luís (1997).

Con el fin de aumentar la tasa de detección de ítems que muestren FDI no uniforme, Mazor et al. (1994) propusieron una modificación del procedimiento estándar de Mantel Haenszel (Holland y Thayer, 1988) que consiste en dividir la muestra total en dos submuestras de alta y baja ejecución y aplicar en cada submuestra el procedimiento de Mantel-Haenszel.

Por otro lado, el procedimiento de análisis de los residuales propuesto consiste en determinar, en primer lugar, el modelo de la TRI que mejor se ajusta a los datos en el grupo tomado como grupo de referencia (muestra de niños), en nuestro caso el modelo logístico de dos parámetros y, una vez evaluado el ajuste, se calcula la curva característica de cada uno de los ítems en ese grupo; posteriormente, mediante el módulo de análisis de residuales del programa GENESTE (San Luís et al., 1995) se calculan los residuos estandarizados del grupo focal (muestra de niñas) con respecto a la curva característica del grupo de referencia, mostrando la representación gráfica de los residuos.

Utilizando los índices de ajuste que ofrece el programa GENESTE —proporción de residuos estandarizados absolutos inferiores a 1,96 o 2,58 (a= 0,05 o 0,01 respectivamente) o χ^2 (Wright & Panchapakesan, 1969)— se considera que un ítem presenta funcionamiento diferencial si no alcanza un apropiado nivel de ajuste.

Ambos procedimientos fueron comparados por Prieto et al. (1996) mediante un estudio de simulación para probar la capacidad de detección de ítems con FDI no uniforme. El FDI no uniforme se produce cuando la probabilidad de responder correctamente a un ítem entre dos grupos no es la misma para todos los niveles de habilidad (Mellenbergh, 1982). Los resultados del citado estudio sugieren que, aunque el procedimiento MH, con la modificación introducida por Mazor et al. (1994) puede ser efectivo para detectar FDI no uniforme, el pro-

cedimiento de análisis de residuales utilizado produce mejores resultados en la detección de ítems con FDI uniforme y no uniforme.

En el presente trabajo se aplican ambos procedimientos de un modo iterativo, es decir, eliminando paso a paso del test aquel ítem que presenta un mayor desajuste y volviendo a aplicar el procedimiento al resto de los ítems hasta conseguir un conjunto depurado de ítems (Camilli y Shepard, 1994). La razón que nos llevó a escoger esta estrategia se basa en que los procedimientos iterativos, como compensación a su más alto costo computacional, presentan una mayor potencia de prueba que cuando son aplicados de una forma no iterativa. (Fidalgo, Mellenbergh y Muñiz, 1998). Sin embargo no es esta su mayor ventaja, sino la reducción del error tipo I que conllevan. Estas ventajas del procedimiento iterativo han sido contrastadas (Fidalgo et al., 1998) en su aplicación al procedimiento estándar de Mantel-Haenszel.

Resultados

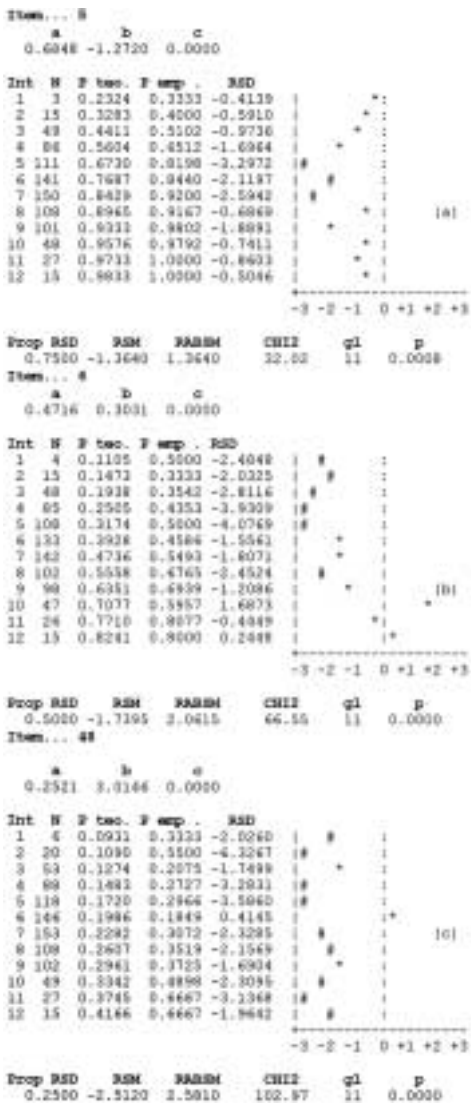
Para obtener un conjunto de ítems libres de FDI fueron necesarias 9 iteraciones al aplicar el procedimiento de MH y 14 iteraciones cuando se utiliza el procedimiento basado en la TRI.

Los ítems que mostraron FDI en cada uno de los procedimientos aparecen a continuación por orden de eliminación:

MH : 34, 48, 5, 6, 7, 10, 37, 3, 43
TRI : 5, 48, 6, 43, 13, 10, 52, 37, 3, 7, 54, 34, 11, 16

Como puede observarse los 9 ítems detectados mediante el procedimiento MH lo son también mediante el procedimiento de la TRI, mientras que los ítems 13, 52, 54, 11 y 16 han sido identificados sólo mediante el procedimiento de la TRI . Asimismo, del total de los 14 ítems diferentes detectados por ambos procedimientos, 7 de ellos corresponden a ítems en los que se detectaron diferencias significativas entre niños y niñas.

La gráfica 1 (a, b y c) muestra la representación gráfica de los residuos estandarizados obtenidos para los ítems 5, 6 y 48, tres de los ítems detectados por ambos procedi-



Gráfica 1: Ítems con FDI uniforme detectados con ambos procedimientos.

mientos en las primeras iteraciones como ítems con FDI.

Los residuos estandarizados comprendidos en el intervalo -2 y +2 corresponderían a aquellos valores en los que no hay diferencias significativas en el rendimiento entre los niños y las niñas, los valores superiores o inferiores a ese intervalo indicarían diferencias significativas entre las curvas características de los niños y las niñas para ese nivel de habilidad.

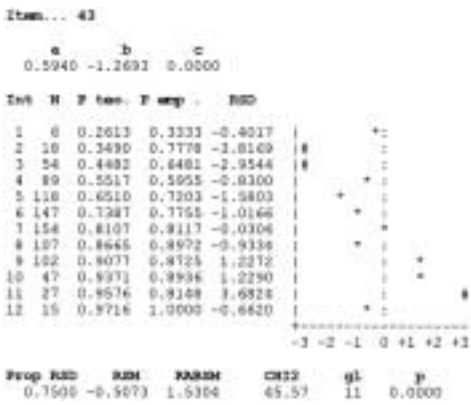
Los valores encontrados a la izquierda del punto cero, indicarían un menor rendimiento en los sujetos del grupo focal (la muestra de niñas) cuando son comparados con los sujetos del grupo de referencia (la muestra de niños), mientras que los valores situados a la derecha del punto cero indicarían un mejor rendimiento en los sujetos del grupo focal.

La gráfica permite analizar si el tipo de FDI encontrado es uniforme o no uniforme. Si todos los residuos estandarizados se encuentran situados en el mismo lado del eje correspondiente al punto cero, el FDI será uniforme; por el contrario, si se encuentran distribuidos a ambos lados del eje será un FDI no uniforme.

Los tres ítems que aparecen en la gráfica han sido más fáciles para el grupo de referencia (muestra de niños) que para el grupo focal (muestra de niñas) y además muestran un FDI uniforme.

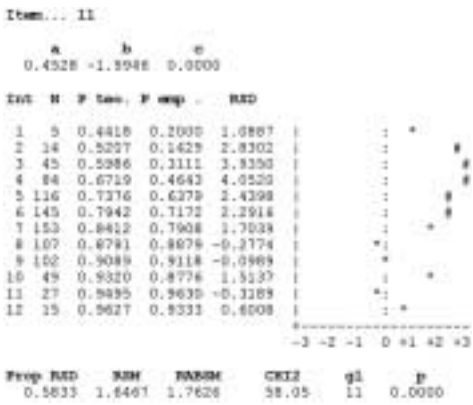
En la gráfica 2 aparece la representación de los residuos estandarizados correspondientes al ítem 43, detectado por ambos procedimientos. Es el único ítem que ha sido detectado por el procedimiento de MH como ítem con FDI no uniforme, mientras que mediante el procedimiento de la TRI se han detectado 4 (13, 43 52 y 54).

Como se puede observar en la gráfica 2, para los niveles más bajos de habilidad el ítem ha sido más fácil para el grupo de referencia (muestra de niños), mientras que para los niveles más altos de habilidad lo ha sido para el grupo focal (muestra de niñas).



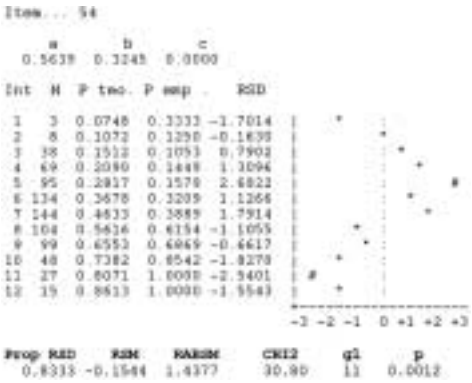
Gráfica 2: Ítems con FDI no uniforme detectado por ambos procedimientos.

En la gráfica 3 se muestran los residuos estandarizados correspondientes al ítem 11, como se puede observar se trata de un ítem que ha sido más fácil para la muestra de niñas, (grupo focal) que para la de niños (grupo de referencia)



Gráfica 3: Ítems de menor dificultad para el grupo focal de las niñas.

La gráfica 4 muestra el ítem 54 que ha sido detectado como ítem con FDI sólo por el procedimiento de la TRI, como se puede observar se trata de un ítem con FDI no uniforme.



Gráfica 4: Ítems con FDI no uniforme detectado sólo mediante TRI.

Discusión

Los resultados de este trabajo nos llevan a considerar la posibilidad de que los procedimientos estadísticos para la detección del FDI pueden verse mejorados con su aplicación de forma iterativa, una estrategia de análisis que ha demostrado en anteriores estudios tanto su mayor potencia como su capacidad para la reducción de la tasa de falsos positivos. Es más, dada la imposibilidad de distinguir entre FDI real y falsos positivos en estudios con datos empíricos (no simulados), la aplicación conjunta de diferentes procedimientos podría llevar a conclusiones más fiables acerca de la identificación de posibles ítems sesgados.

De los dos procedimientos utilizados en este estudio para la detección de ítems que muestren un funcionamiento diferencial, la variación del procedimiento de MH propuesta por Mazor et al. (1994) y el procedimiento de análisis de residuales derivado de las técnicas de la TRI, es éste último el que detecta un mayor número de ítems como potencialmente sesgados, lo que indica la mayor capacidad ya demostrada de este procedimiento para detectar FDI no uniforme (Prieto, Barbero y San Luis, 1997). Nótese que de los 9 ítems detectados por el procedimiento de MH sólo 1 ítem mostraba FDI no uniforme, mientras que de los 5 ítems detectados únicamente mediante el procedimiento derivado de la TRI, eran 3 los que presentaban dicho tipo de FDI. Sin embargo también debe considerarse que el procedimiento de análisis de residuales cuando se ha comparado con las técnicas de MH (Prieto, Barbero y San Luis, 1997) en estudios con datos simulados presenta una mayor tasa de Error Tipo I , aunque ambos procedimientos habían sido aplicados de forma no iterativa.

La imposibilidad que se nos presenta de distinguir en la realidad entre FDI y Error Tipo I es la que nos lleva a abogar por el uso conjunto de diferentes procedimientos de análisis. Así, en nuestro estudio, esta aplicación conjunta de diferentes técnicas nos hace ganar en confianza acerca del posible FDI exhibido por 9 de los 14 ítems detectados en total.

Debemos destacar también la ventaja que ofrece el programa GENESTE de cara a la identificación, de una manera muy sencilla, del tipo de FDI que presentan los ítems y para qué niveles de las diferencias entre la ejecución de los grupos es mayor y, por lo tanto, qué niveles se ven más afectados.

Respecto a la prueba concreta empleada en este estudio, y a la vista de nuestros resultados, concluimos que convendría depurarla antes de ser utilizada para realizar inferencias acerca del nivel de conocimientos de ciencias de los niños y niñas españoles de 13 años y de sus diferencias en el rendimiento ante dicha prueba. Así, de los resultados encontrados parece desprenderse la hipótesis de que algún tipo de factor relacionado con el sexo de los sujetos puede estar influyendo en la forma en la que estos responden a los ítems de la prueba.

Finalmente, estimamos la conveniencia de seguir analizando y perfeccionando los procedimientos para la detección del FDI, por ejemplo, y como se sugiere en este tra-

bajo, estudiando las posibles mejoras que los procedimientos iterativos parecen apuntar, pero sin olvidar la necesidad de ir más allá analizando las fuentes del mismo. Se hace necesario, por ello, la planificación de estudios posteriores para determinar las causas del FDI encontrado en esta prueba, y poder así diferenciar entre *sesgo* o *impacto*.

Referencias

Ackerman, T.A. (1988) An explanation of differential item functioning from a multidimensional perspective. comunicacion presentada en la reunion anual de American Educational research Association, New Orleans.

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensioanl perspective. *Journal of Educational Measurement*, 29(1), 67-91.

Ackerman, T.A. (1993). Differential item functioning as a function od the valid subtest space. Comunicacion presentada en la reunion de la European Psychometric Society. Barcelona.

Angoff, W.H. (1982). Use of difficulty and dsicrimination indices for detecting item bias. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp 96-116). Baltimore: John Hopkins University Press.

Camilli, G., & Shepard, L.A. (1994). *Methods for Identifying Biased Test Items*. London: Sage Pub.

Cole, N.S. & Moss, P.A. (1989). Bias in Test Use. In R.L. Linn (Ed.) *Educational Measurement* (3rd de., pp. 201-219). New York: McMillan.

Fidalgo, A.M. (1996). *Funcionamiento Diferencial de los Items. Procedimiento Mantel-Haenszel y modelos loglineales.* Tesis Doctoral no publicada. Universidad de Oviedo.

Fidalgo, A.M. (1996b). Funcionamiento Diferencial de los Items. En J. Muñiz (Coord.), *Psicometría* (pp. 370-455). Madrid: Universitas.

Fidalgo, A.M., Mellenbergh, G.J. y Muñiz, J. (1998). Comparación del procedimiento Mantel-Haenszel frente a los modelos loglineales en la detección del funcionamiento diferencial de los ítems. *Psicothema*, 10(1), 209-218

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

Jensen, A.R. (1980). *Bias in Mental Testing*. New York. Free Press.

Linn, R.L., & Harsnich, D.L. (1981). Interactions between item content and group membreship on achievement test items. *Journal of Educational Measurement*, 18, 109-118.

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

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

Muñiz, J. (1990): *Teoría de Respuestas a los Items: Un nuevo enfoque en la evaluación psicológica y educativa*.. Pirámide. Madrid.

Oort, F.J. (1992). Using restricted factor analysis to detect item bias. *Methodika*, 6, 150-166.

Oort, F.J. (1993). Theory of violators: Assessing unidimensionality of psychological measures. En R. Steyer, K.F. Wender y K.F. Widaman (Eds.): Psychometric methodology. Proceedinngs of the 7th European Meeting of the Psychometric Society in Trier. Stuttgart: Gustav Fischer Verlag.

Prieto, P., Barbero, M.I., & San Luis, C. (1997). *Identification of nonuniform DIF: A comparison of Mantel-Haenszel and IRT analysis procedure. Educational and Psychological Measurement*, 57, 4: 559 - 568

San Luis, C., Prieto, P., Barbero, M., & Sánchez, J.A. (1995). GENESTE: Un programa de control para TRI. *Psicológica*, 16, 297-304.

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

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

Aceptado el 23 de diciembre de 1998