Revista de Investigación Educativa, 1998, Vol. 16, n.º 1, págs. 187-205

COMPARACIÓN DE MÉTODOS ESTADÍSTICOS PARA LA EVALUACIÓN DEL FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS

Javier Gil Flores, Eduardo García Jiménez* y Gregorio Rodríguez Gómez***

* Área MIDE de la Universidad de Sevilla
** Área MIDE de la Universidad de Cádiz

RESUMEN

En este trabajo llevamos a cabo un estudio comparativo de cuatro métodos para la detección del funcionamiento diferencial de los ítems: el método delta, *uno de los más usados en el contexto de la teoría clásica, los métodos de las* diferencias de proporciones *y de* chi cuadrado sumado, *ambos basados en tablas de contingencia, y el* contraste de parámetros b *obtenidos al estimar la curva característica del ítem. Los datos utilizados corresponden a una prueba de comprensión de textos administrada a 519 alumnos de Educación Secundaria, y la variable diferenciadora de los grupos ha sido el sexo. Tras la aplicación de las diferentes técnicas, obtenemos una cierta convergencia entre los resultados alcanzados, dato que iría en contra de la hipotética superioridad de unos métodos sobre otros.*

ABSTRACT

A comparative study over four methods used to detect the differential item functioning is the focus of this paper. These methods are: the delta *method (frecuently used in the classic theory's context), the methods of* proportion's differences *and the chi square (both based on the contingency's table), and the* contrast of the b parameters, *obtained when stimating the caracteristic curve of the item. The data used has been accomplished from a test about text's comprehension administered to 519 Secondary Education's students, being sex the differencing variable of the groups. After the application of these different techniques, the results obtained are similar.*

INTRODUCCIÓN

Un tema que ha llegado a constituirse en foco de interés para investigadores y constructores de tests es el relativo al sesgo de los instrumentos de medida, tal y como refleja la profusión de trabajos que en los últimos años se han centrado en este tema. De acuerdo con el concepto habitualmente manejado, un instrumento de medida está sesgado cuando no ofrece la misma medida para dos sujetos o grupos de sujetos que cuentan con un nivel similar en el atributo medido, sino que sistemáticamente perjudica a alguno de ellos. Este mismo concepto podría trasladarse a cada uno de los ítems que componen el instrumento. Es decir, un ítem está sesgado cuando sujetos con la misma competencia y pertenecientes a distintas subpoblaciones no cuentan con el mismo grado de acierto al responder el ítem. El problema del sesgo representa un aspecto clave, no sólo desde el punto de vista social, dada la injusticia que representa para los colectivos que se ven perjudicados, sino también desde la óptica psicométrica, por el impacto negativo de los ítems sesgados sobre la validez de la prueba.

Ahora bien, no debe confundirse el sesgo del ítem con diferencias reales en el rendimiento de los grupos. El hecho de que hombres y mujeres, por ejemplo, tengan distinta tasa de aciertos en un ítem no significa que el ítem esté sesgado; puede ocurrir que el ítem discrimine entre ambos grupos por existir entre ellos una diferencia real en la capacidad medida. Al tratar de determinar el sesgo, lo que se persigue será precisamente separar las diferencias reales de las que son generadas por el propio instrumento de medida.

Generalmente, el estudio del sesgo de los ítems se ha vinculado a diferenciaciones de los sujetos desde el punto de vista social, cultural, racial, sexual, religioso, económico, geográfico: negros-blancos, mujeres-hombres, ricos-pobres,... Normalmente se alude a dos grupos, a los que se denomina grupo focal y grupo de referencia, que son respectivamente el grupo minoritario perjudicado o favorecido por el ítem y el grupo de comparación, respecto al cual se produce la diferencia observada en el grupo focal. También sería posible considerar el sesgo comparando grupos de similar tamaño o tomando más de dos grupos.

A veces, el sesgo se debe a la presencia de una variable que contamina la respuesta del individuo. Por ejemplo, si en un test de inteligencia incluimos una pregunta en la que se emplea un lenguaje complejo y elevado, ese elemento estará midiendo no sólo la inteligencia sino la capacidad de comprensión lectora de los sujetos, y resultará sesgado contra los sujetos poco competentes en comprensión lectora. Existen otras múltiples causas de sesgo de los ítems; en realidad, es difícil encontrar pruebas que no estén sesgadas.

La mejor forma de prevenir el sesgo en los ítems consiste en llevar a cabo un cuidadoso análisis del contenido de los mismos por parte de expertos. Sin embargo, cabe la posibilidad de realizar algunos análisis estadísticos mediante los cuales podemos detectar ítems que escaparon al examen previo. Los métodos para evaluar el sesgo se han basado en la comparación de la diferencia de respuesta en los dos grupos registrada para el ítem y la diferencia de respuesta que permanece constante a lo largo del test. Si se asume la unidimensionalidad del test, todos sus ítems miden el mismo cons-

tructo, y por tanto la diferencia en el grado de acierto en dos grupos distintos habría de mantenerse constante a lo largo de todos ellos. Si en un ítem se registran diferencias de acierto que se apartan de las que se dan en la globalidad de la prueba, podremos sospechar que se trata de un ítem con sesgo. Sin embargo, cabría la posibilidad de que todos los ítems contaran con un mismo sesgo y ese ítem destacara sobre los demás. Es decir, se podrían detectar ítems que resultan especialmente sesgados respecto al conjunto de ítems. Por este motivo resulta preferible, desde el punto de vista metodológico, hablar de técnicas para la evaluación del funcionamiento diferencial de los ítems. En la literatura psicométrica se suele recoger este concepto mediante la sigla DIF, que proviene de la expresión inglesa *Differential Item Functioning*.

OBJETIVO DEL ESTUDIO

Existen múltiples métodos estadísticos para evaluar el DIF. Los estudios comparativos y las simulaciones dirigidos a determinar la eficacia de diferentes métodos han sido frecuentes en los últimos años (Cohen y Kim, 1993; Zwick y otros, 1994a, 1994b), y son innumerables las variantes propuestas o las adaptaciones a casos concretos, formuladas a partir de los métodos usados habitualmente (Diamond, 1992; Miller y Spray, 1993; Oshima y otros, 1994; Mazor y otros, 1994, 1995; Nandakumar, 1994; Pang y otros, 1994; Hanson y Feinstein, 1995,...).

A pesar de tales estudios, no existe acuerdo sobre qué métodos resultan más adecuados. Para algunos, los métodos clásicos se ven superados por las nuevas técnicas basadas en la teoría de respuesta a los ítems (TRI), mientras que otros ven en este tipo de procedimientos problemas conceptuales y exigencias teóricas que no siempre se cumplen en la práctica. En realidad, ningún método resulta suficiente por sí mismo. La primera de las directrices ofrecidas por Hambleton y otros (1993) de cara a llevar a cabo estudios sobre el DIF apunta en este sentido, afirmando que no existe un único método capaz de garantizar la detección de todos los ítems de un test afectados por un funcionamiento diferencial.

Ante la diversidad de opiniones, en este trabajo hemos pretendido llevar a cabo una comparación de distintos métodos a fin de comprobar la convergencia o, por el contrario, la discrepancia entre los resultados a que nos conducen. Con este objetivo, aplicaremos diferentes técnicas para la evaluación del DIF a los ítems de una prueba de comprensión de textos, examinando el posible sesgo que éstos presentan en función de la variable sexo. Hemos tomado el factor diferenciador sexo por ser ésta una variable comúnmente considerada en los estudios sobre DIF y por las ventajas prácticas que ofrece, ya que la diferenciación por sexos suele dar lugar a dos subgrupos de examinados de tamaño parecido, evitando que el reducido número de sujetos en alguno de ellos impida la aplicación de algunos métodos de análisis, basados en la TRI, que exigen contar con un mínimo tamaño muestral de cara a la estimación de parámetros.

Por tanto, se tratará de identificar, utilizando diferentes métodos de análisis, los ítems de la prueba que pudieran favorecer o perjudicar sistemáticamente a las alumnas (grupo focal) frente a los alumnos (grupo de referencia) al medir su capacidad para

la comprensión de textos. La comparación de los resultados nos permitirá valorar la coincidencia o discrepancia entre los diferentes métodos usados.

DATOS UTILIZADOS

Los datos utilizados en este estudio han sido extraídos de una investigación que tenía por objetivo la construcción de un test adaptativo computerizado (TAC), cuyo fin era medir la capacidad de los alumnos de Educación Secundaria para comprender información escrita en el área de Geografía e Historia (García y otros, 1993). Aquí hemos tomado una de las 49 pruebas en que fue fragmentado el banco inicial de ítems, de cara a su calibración conforme a un modelo logístico de 3 parámetros. La calibración de un banco de ítems para la construcción de un TAC requiere la aplicación previa de cada una de las pruebas en el formato convencional no adaptativo de lápiz y papel. Los datos obtenidos con una de estas pruebas son los que utilizaremos en la comparación de métodos para la evaluación del DIF.

La prueba consistía en la presentación de diversos textos históricos o relativos a temas de geografía física, humana y económica, que daban paso a la formulación de una serie de cuestiones o ítems sobre su contenido. En cada ítem se ofrecían cuatro opciones de respuesta, de las cuales una sola era correcta, funcionando las tres restantes como distractores. Las puntuaciones correspondientes a cada ítem fueron dicotómicas, asignando 1 en caso de acierto y 0 en caso de error, y la suma de todas ellas se tomó como puntuación alcanzada en la prueba.

El número de examinados ascendió a 519 sujetos, de los cuales 219 eran alumnos y 300 alumnas. Todos ellos cursaban estudios de BUP o niveles equivalentes de la nueva Educación Secundaria, en el I.B.Caballero Bonald, de Jerez de la Frontera, y en el I.E.S. José Luis Tejada, de El Puerto de Santa María, ambos en la provincia de Cádiz. La administración tuvo lugar en el segundo trimestre del curso académico 1994/95. Una vez eliminados, de cara a mejorar la calidad psicométrica de la prueba, los ítems que presentaban una dificultad (o facilidad) extrema y no alcanzaban un nivel mínimo de discriminación (correlación biserial puntual inferior a 0.30), el instrumento quedó constituido por 20 ítems, los cuales han sido objeto del presente estudio.

MÉTODOS PARA LA EVALUACIÓN DEL DIF

La identificación de los ítems afectados por DIF se ha realizado recurriendo a un total de cuatro métodos estadísticos. Hemos seleccionado el método delta, por tratarse de uno de los más ampliamente usados en el contexto de la teoría clásica, y también los métodos de la diferencia de proporciones y de chi cuadrado sumado, ambos basados en el análisis de tablas de contingencia. Por último, hemos considerado una técnica para la identificación del DIF basada en los modelos de la TRI; concretamente, se trata de la comparación estadística de los parámetros *b* obtenidos al estimar la curva característica del ítem para el grupo focal y el grupo de referencia. Tras un breve comentario sobre cada uno de los métodos, presentaremos los resultados obtenidos.

Método delta

Uno de los procedimientos más usados en el contexto de la teoría clásica es el método delta (Angoff, 1982; Angoff y Ford, 1973). El principio básico consiste en encontrar los ítems con mayor discrepancia entre los índices de dificultad calculados para los dos grupos. El método delta supone transformar las proporciones de acierto (índice de dificultad) para los dos grupos en puntuaciones delta (puntuaciones típicas derivadas, con media 13 y desviación típica 4). Por tanto, en primer lugar habrá que transformar los índices de dificultad (proporciones de acierto *p*) en puntuaciones *z* correspondientes al percentil *1-p* de la distribución normal y posteriormente trasladarla a la escala de las puntuaciones delta. Al obtener la puntuación *z* a partir de *1-p* y no a partir de *p*, conseguimos que las puntuaciones altas correspondan a los ítems de mayor dificultad y las puntuaciones bajas a los de menor dificultad. Los índices de dificultad (p_1 y p_2), las puntuaciones z_{1-p} correspondientes a los mismos (z_1 y z_2) y las puntuaciones delta (Δ_1 y Δ_2) para los 20 ítems de la prueba aparecen recogidos en la tabla 1.

Tabla 1

RESULTADOS DE LA APLICACIÓN DEL MÉTODO DELTA

Item	p_1	p_2	z_1	z_2	Δ_1	Δ_2	d
1	0.75	0.66	-0.67	-0.41	10.32	11.36	—.24
2	0.50	0.34	0.00	0.41	13.00	14.64	—.74
3	0.15	0.22	1.04	0.77	17.16	16.08	.66
4	0.86	0.75	-1.08	-0.67	8.68	10.32	—.51
5	0.79	0.74	-0.81	-0.64	9.76	10.44	.01
6	0.79	0.71	-0.81	-0.55	9.76	10.80	—.21
7	0.24	0.21	0.71	0.81	15.84	16.24	—.15
8	0.52	0.43	-0.05	0.18	12.80	13.72	—.30
9	0.34	0.34	0.41	0.41	14.64	14.64	.15
10	0.76	0.76	-0.71	-0.71	10.16	10.16	.39
11	0.59	0.56	-0.23	-0.15	12.08	12.40	.10
12	0.38	0.34	0.31	0.41	14.24	14.64	—.07
13	0.92	0.92	-1.41	-1.41	7.36	7.36	.54
14	0.54	0.52	-0.10	-0.05	12.60	12.80	.14
15	0.48	0.34	0.05	0.41	13.20	14.64	—.63
16	0.36	0.34	0.36	0.41	14.44	14.64	.04
17	0.29	0.28	0.55	0.58	15.20	15.32	.05
18	0.37	0.35	0.33	0.39	14.32	14.56	.02
19	0.23	0.21	0.74	0.81	15.96	16.24	—.09
20	0.58	0.57	-0.20	-0.18	12.20	12.28	.23

Las puntuaciones delta de cada ítem son representadas gráficamente situando en abscisas la puntuación en uno de los grupos, y en ordenadas la puntuación alcanzada en el otro. De este modo, conseguimos un gráfico formado por tantos puntos como ítems posee el test (ver más adelante figura 1). Esta nube de dispersión tendrá la forma de una elipse; cuando los grupos provienen de una misma población, la elipse se estrechará aproximándose a una recta. Si los puntos se encuentran en línea recta, podemos afirmar que los ítems son insesgados, estando sesgados aquéllos que se apartan considerablemente de ella. No obstante, como afirma Muñiz (1992), el hecho de que los ítems se sitúen en línea recta no implica ausencia de sesgo. Podríamos pensar que una línea recta por debajo de la diagonal representa un sesgo contra el grupo representado en ordenadas, mientras que una recta por encima de la diagonal supondría un sesgo contra el grupo representado en abscisas. En realidad, el método delta evalúa la discrepancia entre sesgos, detectando aquellos ítems que se apartan en este rasgo de los restantes que constituyen la prueba.

Se han propuesto índices que permiten caracterizar el sesgo de los ítems sin necesidad de basarnos exclusivamente en la inspección visual del gráfico. Globalmente, el ajuste de los puntos a la recta puede valorarse a partir de la correlación de Pearson entre los dos grupos de puntuaciones delta. En este caso, el valor de la correlación asciende a 0.97, indicando un buen ajuste y por tanto un bajo DIF. Sin embargo, más que una consideración global sobre los 20 ítems de la prueba, nos interesa caracterizar a cada uno de ellos particularmente. Un modo de hacerlo se basa en calcular la distancia de cada punto al eje principal de la elipse, de forma que cuanto más grande sea esta distancia mayor será el sesgo del ítem en relación a los otros. Para calcularla, siguiendo a Angoff y Ford (1973), partimos del eje principal de la elipse, cuya ecuación corresponderá a la recta

$$\Delta_2 = a \cdot \Delta_1 + b$$

donde los valores de las constantes a y b son

$$a = \frac{S^2_{\Lambda_1} - S^2_{\Delta_1} + \sqrt{(S^2_{\Lambda_2} - S^2_{\Lambda_1})^2 + 4 \; r^2_{\Delta_1\Delta_2} \; S^2_{\Lambda_1} \; S^2_{\Lambda_2}}}{2 \; r_{\Delta_1\Delta_2} \; S_{\Delta_1} \; S_{\Delta_2}}$$

$$b = \bar{\Delta}_2 - a \cdot \bar{\Delta}_1$$

El índice de distancia para un ítem j, cuyas puntuaciones delta en ambos grupos son D_{1j} y D_{2j} respectivamente, vendrá expresado por:

$$d = \frac{a \; \Delta_{1j} - \Delta_{2j} + b}{\sqrt{a^2 + 1}}$$

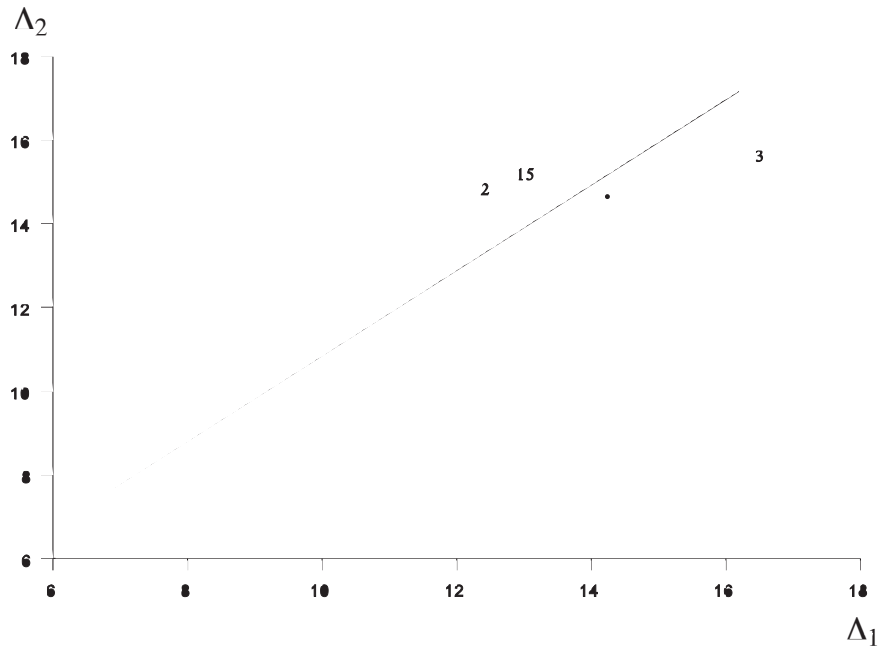
La ecuación para el eje mayor de la elipse que forman los puntos del diagrama de dispersión ha sido determinada en el caso de la prueba que nos ocupa, obteniendo como resultado:

$$\Delta_2 = 0.91 \cdot \Delta_1 + 1.57$$

En la figura 1, este eje ha sido trazado sobre el diagrama de dispersión para las puntuaciones delta, etiquetando por medio de su número a los ítems que más se distancian del eje.

Figura 1

DIAGRAMA DE DISPERSIÓN PARA LAS PUNTUACIONES DELTA, EJE MAYOR DE LA ELIPSE Y PUNTOS MÁS DISTANCIADOS DEL EJE



Calculando la distancia al eje para todos los ítems tenemos una visión global y podemos establecer, siguiendo criterios meramente descriptivos, puntos de corte a partir de los cuales descartar ítems que consideramos especialmente sesgados. Las distancias *(d)* quedaron recogidas en la tabla 1 mostrada anteriormente. Fijando en 0.60 el nivel de corte, los ítems que mayor DIF presentan son, en este orden, los números 2, 3 y 15. El signo positivo asociado a la distancia *d* indica que el ítem favorece al grupo focal, mientras que las distancias negativas corresponden a ítems que lo perjudican.

Una modificación propuesta por Camilli y Sephard (1994) al método de Angoff consiste en realizar una segunda estandarización de los índices de dificultad. En lugar de calcular la distancia de los puntos a la recta, tomamos las puntuaciones $z_{1\text{-}p}$ y las estandarizamos respecto al grupo, a partir de la media y la desviación típica del conjunto de puntuaciones $z_{1\text{-}p}$. Es decir, la nueva puntuación z^* correspondiente a una puntuación z se obtendrá, por tanto, como:

$$z^* = \frac{z - \mu_z}{\sigma_z}$$

donde μ_z es la media de las puntuaciones $z_{1\text{-}p}$ para todos los ítems de la prueba, y σ_z la desviación típica de las mismas. La diferencia $z_1^* - z_2^*$ entre las puntuaciones obtenidas para un mismo ítem en los grupos focal y de referencia puede ser tomada como medida del DIF. No obstante, el cálculo de este valor no altera el resultado obtenido basándonos en la distancia al eje; de nuevo, los valores más altos para la diferencia entre puntuaciones z estandarizadas corresponden a los ítems 2, 3 y 15, donde $z_1^* - z_2^*$ asciende a –0.48, 0.50 y –0,41 respectivamente.

Método de la diferencia de proporciones

Este método, propuesto por Dorans y Kurlick (1983, 1986), se basa en el cálculo de la diferencia de proporciones en cada uno de los niveles en que previamente son subdivididos ambos grupos. Suponiendo que hemos dividido los grupos en k niveles de capacidad, podemos calcular la diferencia en la proporción de aciertos para cada uno de esos niveles. Si p_{2j} y p_{1j} son las proporciones de acierto en los grupos focal y de referencia para el nivel de capacidad j, la diferencia de proporciones vendrá dada por:

$$\Delta p_j = p_{2j} - p_{1j}$$

A partir de la suma ponderada de las diferencias de proporciones en los k niveles, se puede calcular el estadístico:

$$\text{DPE} \quad \text{DIF} = \frac{\sum_{j=1}^{k} w_j \Delta p_j}{\sum_{j=1}^{k} w_j}$$

donde el coeficiente de ponderación w_j puede adoptar diferentes valores, según los propósitos de la investigación (Dorans y Holland, 1993). Se recomienda el uso de n_{2j} o

frecuencia en *j* de los sujetos del grupo focal, pues de esta manera se da más peso a las diferencias de proporciones registradas en niveles de capacidad donde los sujetos del grupo focal son más numerosos.

Los valores del estadístico DPE-DIF están comprendidos entre –1 y 1. Siguiendo las pautas de interpretación ofrecidas por Dorans y Holland (1993), cuando encontramos valores que se sitúan, en valor absoluto, por debajo de 0.05, podemos afirmar la inexistencia de un funcionamiento diferencial de los ítems; valores absolutos comprendidos entre 0.05 y 0.10 no son preocupantes, aunque aconsejan una inspección de los ítems; y valores por encima de 0.10, en valor absoluto, hacen necesario un cuidadoso examen. Los valores positivos indicarán una ventaja para el grupo focal, mientras que los negativos reflejan una desventaja para este grupo.

La aplicación de este método a los ítems de la prueba analizada se ha basado en la división de los grupos en 6 niveles de capacidad. Todos los intervalos creados poseen una amplitud de 3 unidades. La frecuencia de sujetos en cada nivel de dificultad, tanto para el grupo focal como para el grupo de referencia, aparece recogida en la tabla 2.

Tabla 2

NÚMERO DE SUJETOS INCLUIDOS EN CADA NIVEL DE CAPACIDAD

Niveles	G. Referencia	G. Focal
1-3	3	5
4-6	29	48
7-9	63	110
10-12	63	82
13-15	43	37
16-18	18	18

Los valores de las proporciones de acierto en cada nivel, así como el estadístico DPE-DIF calculado para cada uno de los ítems, aparecen en la tabla 3. Considerando afectados de DIF aquellos ítems en los que el estadístico calculado se aproxima o supera en valor absoluto la cota de 0.10, este método señala a los números 2, 3, 4 y 15, de entre los cuales únicamente el ítem 3 favorece al grupo focal, de acuerdo con el signo positivo asignado.

Método de chi cuadrado sumado

Uno de los métodos basado en el estadístico χ^2, es el debido a Camilli (1979). Para cada ítem, se parte de la construcción de tablas referidas a cada nivel de capacidad, en la que se presentan las frecuencias de acierto y error para los grupos focal y de referencia (ver tabla 4).

Tabla 4

TABLA DE CONTINGENCIA PARA EL ÍTEM i Y EL INTERVALO j

		Puntuación en el ítem i		
		Acierto (1)	**Error (0)**	
Grupo	**Referencia**	A_j	B_j	n_{Rj}
	Focal	C_j	D_j	n_{Fj}
		m_{1j}	m_{0j}	T_j

Los valores A_j, B_j, C_j y D_j representan las respectivas frecuencias en el intervalo j-ésimo al que se refiere la tabla. Los valores m_{Aj} y n_{Aj} son frecuencias marginales, y el valor T_j se corresponde con la frecuencia total en el nivel de capacidad considerado. A partir de esta tabla es posible calcular el estadístico χ^2_j, de acuerdo con la siguiente expresión:

$$\chi^2_j = \frac{T_j \left(A_j D_j - B_j C_j\right)^2}{n_{Rj}\, n_{Fj}\, m_{1j}\, m_{0j}}$$

La suma de los valores χ^2_j obtenidos para los k niveles de capacidad se distribuyen según k grados de libertad. Bastará comparar con el correspondiente valor crítico para decidir sobre la hipótesis nula de no existencia de DIF.

Utilizando los mismos intervalos de capacidad considerados al aplicar el método de la diferencia de proporciones, hemos calculado los correspondientes valores χ^2_j, y a partir de éstos el valor de chi-cuadrado sumado de Camilli (χ^2_c). Todos ellos se recogen en la tabla 5. En el caso de ítems muy fáciles o muy difíciles no es posible calcular el valor de χ^2_j en los intervalos de capacidad extremos, dado que existen celdas vacías o con frecuencias esperadas muy bajas. Los ítems en los que afirmamos la existencia de DIF serán aquéllos para los cuales el valor de χ^2_c supera al valor crítico χ^2, que para un nivel de significación α=0.05 adopta los valores 5.99, 7.82 y 9.49 en los casos de 2, 3 y 4 grados de libertad respectivamente. Tales ítems han sido marcados en la tabla 5 mediante la colocación de un asterisco junto al valor de χ^2_c.

Tabla 3

RESULTADOS DE LA APLICACIÓN DEL MÉTODO DE LA DIFERENCIA DE PROPORCIONES

Ítem	Nivel 1		Nivel 2		Nivel 3		Nivel 4		Nivel 5		Nivel 6		$\Sigma(n_J \cdot \Delta_p) / \Sigma n_J$
	p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2	
1	.000	.000	.551	.354	.587	.554	.825	.817	.930	.837	1.000	1.000	—.06
2	.333	.000	.137	.083	.317	.272	.476	.390	.860	.486	1.000	.888	—.11
3	.000	.000	.000	.062	.047	.109	.095	.158	.302	.513	.611	1.000	.10
4	.333	.600	.586	.645	.873	.681	.920	.743	.907	.945	1.000	1.000	—.10
5	.000	.200	.586	.625	.746	.736	.809	.719	.930	.891	.944	1.000	—.02
6	.333	.000	.586	.458	.682	.709	.873	.731	.930	.918	1.000	.888	—.06
7	.000	.000	.137	.020	.111	.163	.158	.195	.348	.405	.888	.722	.01
8	.000	.000	.103	.166	.460	.409	.619	.597	.697	.702	.722	.111	—.05
9	.333	.200	.241	.125	.269	.272	.365	.304	.372	.540	.611	1.000	.01
10	.000	.200	.551	.583	.634	.718	.825	.829	.930	.891	.944	.944	.04
11	.333	.400	.310	.208	.349	.500	.698	.670	.837	.648	.944	1.000	.01
12	.000	.000	.000	.062	.238	.190	.460	.451	.558	.648	.833	.944	.01
13	.333	.400	.827	.708	.873	.936	.968	1.000	.976	1.000	1.000	1.000	.02
14	.000	.200	.103	.250	.396	.390	.634	.646	.767	.756	.888	1.000	.03
15	.000	.000	.172	.104	.190	.209	.666	.390	.697	.621	.888	.944	—.09
16	.333	.000	.137	.270	.254	.236	.349	.353	.534	.621	.722	.666	.02
17	.000	.000	.069	.125	.190	.154	.222	.317	.465	.486	.833	.777	.02
18	.333	.200	.103	.208	.381	.254	.365	.402	.418	.405	.611	.944	.00
19	.333	.000	.103	.166	.095	.127	.254	.207	.279	.189	.611	.833	.01
20	.000	.200	.241	.187	.381	.381	.650	.743	.883	.973	.944	1.000	.03

Tabla 5

RESULTADOS DE LA APLICACIÓN DEL MÉTODO DE CHI CUADRADO SUMADO

Ítem	χ^2_2	χ^2_3	χ^2_4	χ^2_5	χ^2_c
1	2.881	.175	.016	.	3.07
2	.	.390	1.075	12.946	14.42 *
3	.	1.911	1.253	3.695	6.86
4	.273	7.840	7.561	.	15.67 *
5	.114	0.019	1.576	.	1.71
6	1.182	.134	4.335	.	5.65
7	.	.893	.320	.271	1.48
8	.590	.429	.068	.002	1.09
9	1.744	.001	.583	2.280	4.61
10	.073	1.293	.003	.	1.37
11	1.012	3.687	.126	3.771	8.60
12	.	.541	.011	.678	1.23
13	1.383	2.029	.	.	3.41
14	2.475	.005	.020	.012	2.51
15	.745	.086	10.893	.514	12.24 *
16	1.856	.067	.003	.612	2.54
17	.	.370	1.604	.036	2.01
18	1.417	3.044	.209	.014	4.69
19	.590	.402	.441	.887	2.32
20	.318	.000	1.480	.	1.80

Método del contraste de parámetros

Los métodos de estudio del sesgo basados en la TRI son muy populares entre los investigadores. La idea de partida es que si la probabilidad de acertar un ítem es similar para los sujetos del grupo focal y del grupo de referencia, entonces las respectivas curvas características del ítem en ambos grupos han de ser también similares. En consecuencia, no deben existir diferencias significativas entre las áreas delimitadas por las correspondientes CCI o entre los parámetros que describen a las curvas para los grupos focal y de referencia. Los métodos para comprobar la significación estadística de las diferencias entre parámetros son variados. Aquí llevaremos a cabo un contraste estadístico para la diferencia de parámetros *b* estimados en cada grupo, siguiendo el procedimiento que describimos a continuación.

Si denominamos b_F al parámetro *b* para el grupo focal y b_R al parámetro *b* para el grupo de referencia, someteremos a contraste la hipótesis nula de igualdad frente a la hipótesis alternativa que afirma la existencia de diferencias.

$$H_0 : \Delta b = b_F - b_R = 0$$
$$H_1 : \Delta b = b_F - b_R \neq 0$$

El estadístico de contraste utilizado (Lord, 1980) es el cociente entre la diferencia de parámetros y el error típico para las diferencias. Teniendo en cuenta que este error típico es

$$S_{\Delta b} = \sqrt{S^2_{b_F} + S^2_{b_R}}$$

el estadístico de contraste será:

$$z_b = \frac{b_F - b_R}{\sqrt{S^2_{b_F} + S^2_{b_R}}}$$

que se distribuye normalmente N(0,1). La consulta de una tabla de valores para la distribución normal nos permitirá conocer la significación estadística en este contraste.

En esta prueba se requiere que los parámetros estimados en ambos grupos se encuentren en una misma métrica, antes de que la comparación se lleve a efecto. El procedimiento utilizado para obtener el valor z_b a partir de parámetros b_F y b_R expresados en una misma métrica ha sido el siguiente:

a) Calibrar los ítems en el grupo de referencia, asumiendo un modelo logístico de 3 parámetros, y determinar los parámetros *b* y el error típico de los mismos. En la estimación de parámetros hemos adoptado el método de máxima verosimilitud marginal, debido a Bock y Lieberman (1970), e implementado mediante el programa XCALIBRE (Assessment Systems Corporation, 1995).
b) Calibrar los ítems en el grupo focal, tomando los parámetros del grupo de referencia y sus errores típicos como definición de la métrica común. El método para la equiparación de parámetros ha sido el de la media y la desviación típica. Estudios basados en simulación han comprobado que, para muestras suficientemente amplias, el método de la media y la desviación típica y el método basado en la curva característica del test conducen a resultados similares en la posterior detección del DIF (Kim y Cohen, 1992).
c) Calcular la medida DIF como diferencia entre los parámetros *b* de ambos grupos y valorar la significación de la diferencia.

La tabla 6 muestra los valores obtenidos para los parámetros b en los grupos focal y referencia, expresados en una métrica común, así como los errores típicos para los mismos. La última columna de esta tabla recoge el valor del estadístico de contraste $Z_{b.}$ Como puede apreciarse, fijando un nivel de significación α=0.01 (valores críticos $z_{\alpha/2}$=± 2.57) únicamente la diferencia correspondiente al ítem 2 resulta estadísticamente significativa. No obstante, si hubiéramos sido menos restrictivos (α=0.05 y valores críticos $z_{\alpha/2}$=± 1.96) las diferencias de parámetros para ítems como el 3 y el 15, cuyo DIF fue detectado por métodos anteriores, también habrían resultado significativas. Los signos unidos a los valores del estadístico de contraste resultan coincidentes con los que adoptaban indicadores utilizados en otros métodos.

Tabla 6

RESULTADOS DE LA APLICACIÓN DEL MÉTODO DE LA DIFERENCIA DE PARÁMETROS

Ítem	b_F	b_R	S_{bR}	S_{bR}	Z_b
1	-0.59	-0.72	.141	.163	—.60
2	1.47	0.65	.222	.163	-2.98
3	1.68	2.58	.220	.380	2.05
4	-1.13	-1.56	.164	.186	-1.73
5	-1.15	-1.04	.173	.180	.44
6	-0.89	-1.04	.157	.167	—.65
7	2.29	2.00	.313	.266	—.71
8	1.53	0.67	.284	.204	-2.46
9	1.50	2.23	.222	.342	1.79
10	-1.24	-0.80	.159	.158	1.96
11	0.17	0.15	.172	.150	—.09
12	1.17	1.25	.184	.203	.29
13	-2.48	-2.04	.193	.203	1.57
14	0.27	0.39	.152	.160	.54
15	1.26	0.67	.195	.166	-2.30
16	1.75	1.71	.263	.254	—.11
17	1.96	1.88	.275	.260	—.21
18	1.58	1.90	.238	.289	.85
19	2.37	2.37	.321	.343	.00
20	-0.11	0.17	.123	.151	1.44

COMPARACIÓN DE RESULTADOS

Tras la aplicación de los diferentes métodos para la identificación de ítems que presentan DIF, vamos a calibrar el grado de acuerdo existente entre los resultados logrados con cada uno de ellos. La tabla 7 presenta una síntesis comparativa de las técnicas aplicadas, indicando que todos los ítems detectados lo son al menos por dos de las cuatro técnicas aplicadas. Además, los cuatro procedimientos de análisis han señalado el ítem 2 como ítem afectado de DIF y existe también amplio acuerdo en torno al ítem 15. A esto podemos sumar la coincidencia en el signo del sesgo, en los casos en que éste es considerado. Así, las tres técnicas que informan sobre el sentido del sesgo (métodos delta, de la diferencia de proporciones y del contraste de parámetros) convergen al indicar que los ítems 2, 4 y 15 perjudican a los examinados pertenecientes al grupo focal, mientras que el ítem 3 los favorece.

Los criterios establecidos para considerar un ítem sesgado pueden llegar a ser un tanto arbitrarios. En el caso del método delta, hemos fijado una distancia mínima de 0.60 al eje mayor de la elipse, aunque no existe ningún motivo que nos impidiera haber tomar un nivel inferior o superior como punto de corte. Los métodos de chi cuadrado sumado y del contraste de parámetros requieren que el analista fije un nivel de significación para decidir cuándo un ítem presenta DIF. También el método de la diferencia de proporciones deja un tanto abiertas las pautas o criterios de decisión. Por estas razones, más que etiquetar a los ítems como afectados o no afectados por un funcionamiento diferencial, consideramos adecuado hablar del grado en que presentan este rasgo.

Tabla 7

ÍTEMS QUE PRESENTAN FUNCIONAMIENTO DIFERENCIAL, DE ACUERDO CON LOS CUATRO MÉTODOS UTILIZADOS

Método	Ítem			
	2	3	4	15
Método delta (Angoff y Ford, 1973)	*	*		*
Diferencia de proporciones (Dorans y Kurlick, 1983,1986)	*	*	*	*
Chi-cuadrado sumado (Camilli, 1979)	*		*	*
Contraste de parámetros b (Lord, 1980)	*			

Cada uno de los métodos usados proporciona índices que pueden ser tomados como medidas del grado en que el ítem está afectado por DIF: la distancia al eje de la elipse (método delta), el estadístico DPE-DIF (método de la diferencia de proporciones), el valor medio que alcanza c^2_j en los distintos tramos de capacidad en que son divididos los grupos (método de chi cuadrado sumado) o el valor z_b (método del contraste de parámetros). En consecuencia, un estudio de la convergencia entre los distintos métodos podría apoyarse en la correlación que se da entre los respectivos indicadores de DIF, tomados todos ellos en valores absolutos. De este modo, podríamos obtener una medida de la convergencia entre métodos incluyendo en su cálculo a todos y cada uno de los 20 ítems que constituyen la prueba. La matriz de correlaciones intermétodos aparece en la tabla 8, reflejando una moderada o incluso alta convergencia entre los resultados a los que nos conduce cada método.

Tabla 8

CORRELACIONES ENTRE LOS VALORES DEL DIF ESTIMADOS POR DIFERENTES MÉTODOS

	Método delta	**Diferencia de proporciones**	**Chi-cuadrado**	**Contraste de parámetros**
Método delta	1.0000			
Diferencia de proporciones	.8455	1.0000		
Chi-cuadrado	.7129	.7661	1.0000	
Contraste de parámetros	.8561	.8561	.5147	1.0000

Los valores de la correlación de Pearson indican que las medidas de DIF obtenidas a partir del método delta son las que mayor covariación registran con las resultantes de los demás métodos, obteniéndose correlaciones de 0.8561, 0.8455 y 0.7129 con los métodos del contraste de parámetros, de la diferencia de proporciones y de chi cuadrado respectivamente. La correlación más elevada se ha registrado entre el método delta y el método de contraste de parámetros *b*. La coincidencia entre los resultados alcanzados por uno y otro método se confirma si tenemos en cuenta que el método delta destacaba a los ítems 2, 3 y 15 como los que registraban mayor DIF, y estos tres elementos se encuentran también entre los cuatro ítems que presentan una mayor diferencia de parámetros *b*.

CONCLUSIÓN

El desarrollo de los métodos para la evaluación del DIF ha surgido como respuesta a la necesidad de identificar ítems que presentan sesgo. No obstante, las técnicas ideadas cuentan con limitaciones que debemos tener presentes. En el caso del método delta, una crítica habitual es la posible confusión entre dificultad y discriminación. La dificultad diferencial no permanecerá constante a lo largo de los ítems de una prueba, a menos que todos presenten un mismo nivel de discriminación. Es posible que los ítems identificados por un funcionamiento diferencial se caractericen en realidad por un alto poder de discriminación. Por otra parte, el que todos los puntos representados en el gráfico se alineen supone ausencia de DIF, pero ello no implica ausencia de sesgo; cabría la posibilidad de que todos los ítems estén igualmente sesgados. Al aplicar este método tendríamos que partir del supuesto de que la mayoría de los ítems son carentes de sesgo. En general, los métodos basados en la teoría clásica, presentan un problema circular: para la valoración del sesgo es preciso partir de diferentes niveles de capacidad medidos por el propio test, pero si éste presenta sesgo, tales niveles no serían una referencia válida.

Los métodos basados en la TRI salvan este problema, al apoyarse en la comparación de las curvas características del ítem estimadas para cada subgrupo, en lugar de tomar índices de dificultad o proporciones de aciertos. Sin embargo, y a pesar de que se ha demostrado la utilidad práctica de estos métodos, existe en ellos una contradicción teórica: si se asume que la estimación de parámetros es invariante respecto de los sujetos, resultaría imposible que dos submuestras generasen dos curvas diferentes. Como afirman Muñiz y Hambleton (1992), la realidad es más boscosa que la teoría y lo cierto es que, a pesar de unas bases teóricas no todo lo sólidas que se desearía, estos métodos detectan con buena precisión los ítems sesgados.

Tal vez la discusión sobre la superioridad de unos métodos u otros pueda resultar estéril a la luz de resultados como los obtenidos en este trabajo. El alto grado de coincidencia en los valores a los que nos conducen los métodos para la evaluación del DIF bien nos permitiría su utilización indistinta. En particular, en el caso aquí estudiado, la convergencia entre las medidas DIF obtenidas mediante el método delta (el más usado en el contexto clásico) y el contraste de parámetros *b* ponen de manifiesto que los planteamientos apoyados en la teoría clásica y en la TRI no son, en la práctica, tan alejados como pudiera parecer. Con ligeras discrepancias, los diferentes métodos nos llevan a soluciones bastante similares, y de confirmarse este tipo de resultados en el análisis de otros casos, la elección de uno u otro podría basarse más en criterios de operatividad o preferencias personales que en la pretendida superioridad de algunos de ellos.

Por ejemplo, al elegir los métodos basados en la TRI, podríamos tener presentes los inconvenientes prácticos que se asocian a éstos, pues precisan muestras relativamente amplias para conseguir una suficiente bondad de ajuste en la estimación de parámetros, y requieren de cálculos complejos y procedimientos sofisticados. Ambas razones podrían limitar su utilización frente a los restantes métodos, más intuitivos y con menos exigencias en cuanto a tamaño muestral o esfuerzo de cálculo.

A pesar de todo, aunque los métodos coincidan en señalar cuáles son los ítems que presentan DIF, es difícil interpretar las razones del mismo. Todos estos métodos cuentan con el inconveniente de no ofrecer pistas para la interpretación de las causas que originan el posible sesgo. De ahí que un posterior análisis lógico de la redacción y el contenido del ítem sean imprescindibles de cara a mejorar el reactivo o decidir su exclusión del instrumento.

REFERENCIAS BIBLIOGRÁFICAS

Angoff, W.H. (1982). Uses of difficulty and discrimination indices for detecting item bias, en R.A. Berk (Ed.) *Handbook of methods for detecting item bias*. Baltimore, John Hopkins University Press.

Angoff, W.H. y Ford, S.F. (1973). Item-race interaction on a test of scholaastic aptitude, en *Journal of Educational Measurement*, 10, 95-105.

Assesment Systems Corporations (1995). *XCALIBRE. Marginal maximum likelihood estimation program for the 2-and 3-parameter IRT model. Version 1.00*. St. Paul, MN: Autor.

Bock, R.D. y Lieberman, M. (1970). Fitting a response model to n dichotomously scored items, en *Psychometrika*, 35, 179-197.

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

Camilli, G. y Sephard, L.A. (1994). *Methods for identifying biased test items*. Thousand Oaks, California, Sage Publications.

Cohen, A.S. y Kim, S.H. (1993). A comparison of Lord's chi-square and Raju's area measures in detection of DIF, en *Applied Psychological Measurement*, 17, 39-52.

Diamond, J.J. (1992). A Graphic Procedure for Studying Differential Item Functioning, en *Journal of Experimental Education,* 60 (4), 351-57.

Dorans, N.J. y Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. En P.W. Holland y H. Wainer (Eds.), *Differential item functioning*. Hillsdale, New Jersey, Lawrence Erlbaum.

Dorans, M.J. y Kurlick, E.M. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: an application of standardization approach. Princeton, New Jersey, Educational Testing Service.*

Dorans, M.J. y Kurlick, E.M. (1986). Demonstrating the utility of the standardization approach to assessing unespected differential item performance on the Scholastic Aptitude Test, en *Journal of Educational Measurement*, 23, 355-368.

García, E. (Dir.) (1993). *Los Tests Adaptativos Computerizados (TACs): su aplicación al Area de Geografía, Historia y Ciencias Sociales en la Educación Secundaria Obligatoria*. Proyecto aprobado por el CIDE en el Concurso Nacional de Proyectos de Investigación. Convocatoria 1993.

Hambleton, R.K. y otros (1993). Advances in the detection of differentially functioning test items, en *European Journal of Psychological Assessment*, 9 (1), 1-18.

Hanson, B.A. y Feinstein, Z.S. (1995). A Polynomial Loglinear Model for Assessing Differential Item Functioning. Comunicación presentada al *Annual Meeting of the American Educational Research Association.* San Francisco, CA.

Kim, S.H. y Cohen, A.S. (1992). Effects of linking methods on detection of DIF, en *Journal of Educational Measurement*, 29(1) 51-66.

Lord (1980). *Applications of item response theory to practical testing problems*. Hillsdale, Nueva Jersey, LEA.

Mazor, K.M. y otros (1994). Identification of Nonuniform Differential Item Functioning Using a Variation of the Mantel-Haenszel Procedure, en *Educational and Psychological Measurement*, 54 (2), 284-91.

Mazor, K.M. y otros (1995). Using Logistic Regression and the Mantel-Haenszel with Multiple Ability Estimates to Detect Differential Item Functioning, en *Journal of Educational Measurement*, 32 (2), 131-144.

Miller, T.R. y Spray, J.A. (1993). Logistic Discriminant Function Analysis for DIF Identification of Polytomously Scored Items, en *Journal of Educational Measurement*, 30 (2), 107-22.

Muñiz, J. (1992). *Teoría clásica de los tests*. Madrid, Pirámide.

Muñiz, J. y Hambleton, R.K. (1992). Medio siglo de Teoría de Respuesta a los Ítems, en *Anuario de Psicología*, 52, 41-66.

Nandakumar, R. (1994). Development of a Valid Subtest for Assessment of DIF/Bias. Comunicación presentada al *Annual Meeting of the American Educational Research Association*. New Orleans, LA.

Oshima, T.C. y otros (1994). Differential Item Functioning for a Test with a Cutoff Score: Use of Limited Closed-Interval Measures, en *Applied Measurement in Education*, 7 (3), 195-209.

Pang, X. y otros (1994). Performance of Mantel-Haenszel and Logistic Regression DIF Procedures over Replications Using Real Data. Comunicación presentada al *Annual Meeting of the American Educational Research Association*. New Orleans, LA.

Zwick y otros (1994a). Assessment of Differential Item Functioning for Performance Tasks, en *Journal of Educational Measurement*; 30 (3), 233-51.

Zwick, R. y otros (1994b). A Simulation Study of Methods for Assessing Differential Item Functioning in Computerized Adaptive Tests, en *Applied Psychological Measurement*, 18 (2), 121-140.