Psicológica (1998) 19, 201-215.

Diferencias instruccionales y funcionamiento diferencial de los ítems: Acuerdo entre el método Mantel-Haenszel y la regresión logística.

José Luis Padilla[*], Andrés González y Cristino Pérez

Universidad de Granada

Durante las dos últimas décadas, la investigación sobre el sesgo en los tests ha estado centrada en el desarrollo de métodos estadísticos adecuados para detectar ítems con un funcionamiento diferencial. La comprensión del funcionamiento diferencial de los ítems (DIF) no ha recibido tanta atención. Este estudio investiga el efecto de las diferencias instruccionales sobre el DIF. Se utilizó un diseño experimental para inducir DIF manipulando la instrucción que recibían dos grupos de personas. El estudio también comparó la capacidad para detectar DIF de los métodos estadísticos χ^2 de Mantel-Haenszel (Holland y Thayer, 1988) y la regresión logística (Swaminathan y Rogers, 1990). El procedimiento experimental pretendía producir DIF en 9 ítems. Los dos métodos estadísticos identificaron los 9 ítems con DIF previsto.

Palabras clave Funcionamiento diferencial del ítem, causas, Mantel-Haenszel, regresión logística.

Las investigaciones sobre el funcionamiento diferencial de los ítems (DIF) han estado centradas en el desarrollo de métodos estadísticos para identificar de forma fiable aquellos ítems que reflejan una ejecución diferencial de personas igualmente capaces, pero miembros de diferentes grupos demográficos. Por el contrario, la comprensión de las causas del DIF no ha recibido una atención semejante (Scheuneman, 1982, 1987; Skagg y Lissitz, 1992; Schmitt, Holland y Dorans, 1993).

[*] Dirigir la correspondencia a José Luis Padilla. Dept. Psicología Social y Metodología. Facultad de Psicología. Universidad de Granada. Campus de Cartuja. 18071 Granada. Telf: 958 24 62 69. Fax: 958 24 37 46. E-mail: jpadilla@platon.ugr.es

Los pocos resultados obtenidos se pueden resumir en: (1) el efecto de las características superficiales de los ítems se puede explicar recurriendo a diferencias en las experiencias instruccionales de los grupos de personas (Angoff y Ford, 1973; Linn y Harnish, 1981; O'Neill y McPeek, 1993; Scheuneman y Gerritz, 1990; Schmitt y Dorans, 1990); (2) las variables demográficas son etiquetas "muy gruesas" que pueden esconder variables instruccionales relevantes para explicar el DIF (Miller y Linn,1988; Muthén, 1988; Tatsuoka, Linn, Tatsuoka y Yamamoto, 1988); y (3) los métodos estadísticos pueden detectar el DIF cuando se comparan grupos definidos por sus experiencias instruccionales (Padilla, Pérez y González, 1998).

Las razones que pueden explicar la escasez de resultados significativos son: (1) pocos estudios han investigado las causas del DIF; (2) los tests analizados suelen ser tests comerciales, por lo que es raro encontrar ítems con un DIF significativo; y (3) pocas investigaciones han inducido DIF experimentalmente.

La utilización de una aproximación experimental para comprender el DIF ha sido frecuentemente recomendada (Mellenbergh, 1989; Scheuneman, 1987; Schemeiser, 1982; Schmitt, Holland y Dorans, 1993). También el DIF ha sido inducido en investigaciones sobre las características de los métodos estadísticos (Kok, Mellenbergh y Van der Flier, 1985).

Este estudio indujo DIF manipulando diferencialmente la instrucción recibida por dos grupos de personas. El objetivo del estudio era evaluar si diferencias en la instrucción están asociadas con el DIF. Además, se analizó el acuerdo en la detección del DIF inducido entre los métodos estadísticos χ^2 de Mantel-Haenszel (Holland y Thayer, 1988) y la regresión logística (Swaminathan y Rogers, 1990).

METODO

Participantes y diseño La muestra estaba formada por 324 personas, de ellas 241 (74.38%) eran mujeres, y 83 hombres (25.62%). La mediana de la edad era de 22 años. Todas cursaban la asignatura de Psicometría dentro del tercer curso de la Licenciatura de Psicología. El área de contenido elegida para la manipulación instruccional fue "Introducción a la Teoría de la Generalizabilidad" (TG). Ninguna de las personas de la muestra había estudiado antes el área de contenido.

Las personas fueron asignadas al azar a dos grupos: 173 personas al Grupo de Referencia (GR), y 151 al Grupo Focal (GF).

Diferencias instruccionales. Numerosos estudios muestran que, durante el aprendizaje, las personas elaboran representaciones –"modelos mentales"– que dirigen su ejecución en tareas de evaluación. Por ejemplo, Zorroza y Sánchez-Cánovas (1995) mostraron la importancia de los modelos para resolver problemas matemáticos. La enseñanza dirigida a la adquisición de un modelo mental utiliza diagramas, ejemplos y "no-ejemplos" (problemas en los que no se puede aplicar el principio o procedimiento representado en el diagrama).

La manipulación instruccional consistió en seguir en el GR una enseñanza dirigida a la adquisición de un modelo mental sobre un apartado del área de contenido, mientras que en el GF se seguía una enseñanza "tradicional" (meramente descriptiva) para el mismo apartado.

Variable instruccional. Las diferencias instruccionales fueron plasmadas en diferentes unidades de tratamientos. Las unidades de tratamiento son informes escritos elaborados por los autores que presentan la misma información sobre el tema TG. Los informes que recibían los dos grupos diferían en el modo de presentación. Estas diferencias se limitaban al apartado del tema: "Interpretación de los componentes de varianza estimados", ya que este era el apartado sobre el que se deseaba realizar una estrategia instruccional diferencial (EID). La Tabla 1 presenta el esquema del proceso instruccional que se seguía con los dos grupos en el apartado objeto de una EID.

Tabla 1. Proceso Instruccional.

GRUPOS	ESTRATEGIA INSTRUCCIONAL	MODO DE PRESENTACION	NUMERO DE ÍTEMS
Grupo de Referencia	Modelo mental	1 diagrama + 9 ejemplos + 4 no-ejemplos	9 ítems con EID
Grupo Focal	Presentación descriptiva	4 ejemplos	

El diagrama presentado al GR representaba un modelo que describía la utilización de los componentes de varianza estimados, para optimizar el diseño de los estudios de decisión. Los ejemplos y "no-ejemplos" interpretaban los resultados de un ANOVA aplicado a los datos de un estudio de generalizabilidad. La interpretación recomendaba aumentar el número de facetas o analizar los residuales en función de los componentes de varianza estimados. La secuencia de presentación de los contenidos fue la misma en los dos informes.

La elaboración de los informes se hizo por los autores de acuerdo con el contenido sobre la TG que aparece en los manuales de Psicometría (Crocker y Algina, 1986).

Instrumentos de medida. El instrumento de medida fue un test de rendimiento elaborado para medir la ejecución de las personas en el tema TG. Estaba formado por 50 ítems de elección múltiple con 3 alternativas de respuesta. El sistema de puntuación de las respuestas era dicotómico. El test contenía 9 ítems diseñados para medir el contenido instruccional objeto de una EID. Los ítems con EID demandaban la interpretación de los componentes de varianza estimados, de la forma mostrada por los ejemplos y "no-ejemplos" presentados en las unidades de tratamiento.

La fiabilidad del test estimada con el coeficiente alfa fue elevada (.81) a pesar de que contribuyen los ítems diseñados para mostrar DIF.

Procedimiento. El estudio del contenido de los informes y la administración del test de rendimiento se realizó en sesiones de grupo. El tiempo para el estudio de los informes y la administración del test fue estrictamente controlado. Después de estudiar el material respondían al test para lo que disponían de una hora y media.

Técnicas estadísticas

Dimtest. El procedimiento DIMTEST fue desarrollado por Stout (1987) para determinar si un conjunto dado de respuestas a ítems dicotómicos cumple el supuesto de unidimensionalidad esencial. La unidimensionalidad esencial significa que los items miden principalmente la misma habilidad dominante pero que algún ítem puede medir también otra habilidad. A continuación, presentamos de forma breve la lógica del procedimiento.

El usurario debe dividir los ítems en dos tipos de subtests: subtest de evaluación (en la nomenclatura del programa "AT1") y subtest de igualación ("PT"). Los ítems de ambos subtests deben ser dimensionalmente distintos. Los ítems de AT1 deben medir la misma habilidad dominante, mientras que los ítems de PT medirán también esa misma habilidad sólo si se cumple el supuesto de unidimensionalidad esencial.

La elección de los ítems para AT1 se puede realizar mediante el juicio de expertos (fijada por el usuario), o por métodos de análisis exploratorio de datos como el análisis factorial (elección automática por el programa). Si el usuario elige los ítems, hasta un cuarto del total de ítems puede formar AT1; si se opta

por la elección automática, el programa selecciona los ítems con las cargas factoriales más elevadas en el segundo factor antes de la rotación. El subtest PT es utilizado para dividir a los sujetos en *k*-subgrupos con la misma puntuación total.

La expresión matemática del estadístico *T* de Stout es la siguiente:

$$T = \frac{1}{K^{1/2}} \sum_{k=1}^{K} \left[\frac{\hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2}{S_k} \right] \qquad (1)$$

donde:

$\hat{\sigma}_k^2$ = estimación usual de la varianza

= estimación "unidimensional de la varianza (i.e., varianza de n variables de Bernoulli)

S_k= error estandar de estimación para el subgrupo *k*

El estadístico *T* de Stout es la diferencia estandarizada entre dos estimaciones de la varianza: la estimación de la varianza observada real y la estimación unidimensional para cada grupo con la misma puntuación total en AT1. Si el supuesto de unidimensionalidad esencial se cumple, ambas estimaciones de la varianza serán iguales, pero si el test es multidimensional, la estimación de la varianza observada resultará inflada.

Numerosos estudios han analizado la utilidad de DIMTEST para evaluar la unidimensionalidad esencial (Nandakumar, 1991, 1994). Recientemente, Hattie, Krakowki, Roger y Swaminathan (1996) concluyeron que DIMTEST detectaba de manera eficiente desviaciones de la unidimensionalidad cuando el procedimiento formaba automáticamente el subtest AT1; y Padilla, Pérez y González (en prensa) han mostrado su utilidad para examinar el efecto de la instrucción sobre la dimensionalidad de las respuestas a los ítems de rendimiento.

Regresión logística. Swaminathan y Rogers (1990) propusieron el procedimiento de regresión logística (RL) para identificar tanto el DIF uniforme como el no uniforme en ítems dicotómicos. La expresión del modelo de RL es la siguiente:

donde:

siendo *u* la respuesta al ítem, θ el nivel de habilidad de las personas, *g* el grupo de pertenencia (GR o GF), y θg el producto de las variables independientes θ y *g*. El parámetro τ_1 representa la diferencia en habilidad (θ), el parámetro τ_2 las diferencias entre los grupos en la ejecución en el ítem, y el parámetro τ_3, la interacción entre la pertenencia grupal y el nivel de habilidad. Según el modelo, un ítem muestra DIF uniforme, sí τ_2 es distinto de cero y τ_3 es igual a cero; y DIF no uniforme, si τ_3 es distinto de cero con independencia del valor que adopte τ_2. Se ha utilizado el estadístico de Wald que compara el parámetro estimado con su error estandar para examinar la hipótesis de DIF para estos ítems. El análisis para la aplicación de la regresión logística se realizó mediante el programa correspondiente del paquete SPSS (SPSS, 1993).

Mantel-Haenszel. La prueba de Mantel-Haenszel (MH) fue aplicada al análisis del DIF por Holland y Thayer (1988). La formulación de las medidas de DIF que proporciona el procedimiento se puede encontrar en numerosas referencias (Navas y Gómez, 1994). Los dos aspectos del procedimiento más relevantes para su aplicación en este estudio son: la determinación del número de niveles en el criterio de igualación y la posibilidad de detectar DIF no uniforme.

La igualación de la habilidad de las personas en el método MH se ha realizado a partir de los quintiles de la distribución conjunta de puntuaciones totales (igualación gruesa), en lugar de la habitual a partir de las puntuaciones totales individuales (igualación delgada). Tres son las razones de esta decisión: (1) obtener mayor estabilidad en las estimaciones de las frecuencias esperadas; (2) utilizar la mayor parte de los datos disponibles, reduciendo el número de filas y columnas con frecuencia cero; y (3) contar con el mayor número posible de categorías para la habilidad (Fidalgo, 1996). Diversos estudios han mostrado que la estrategia de igualación gruesa proporciona estimaciones precisas de los índices de DIF (Raju, Bode y Larsen, 1989), y los mejores resultados cuando la medida de DIF es el estadístico MH-χ^2 (Donoghue y Allen, 1993). Por otra parte, Hambleton, Clauser, Mazor y Jones (1993) mostraron que no hay diferencias entre los resultados de las diferentes estrategias de igualación, si las distribuciones de habilidad son semejantes.

Numerosos estudios han señalado la incapacidad del método MH para detectar DIF no uniforme (Swaminathan y Rogers, 1990). La modificación

propuesta por Mazor, Clauser y Hambleton (1994) ha sido utilizada en este estudio para mejorar la interpretación de los resultados al comparar el método MH con la técnica RL.

Los valores de los estadísticos para el método MH fueron obtenidos con un programa elaborado por los autores.

Purificación del criterio de igualación. La purificación del criterio de igualación es una práctica aceptada para evitar el problema de la circularidad en la detección del DIF. El criterio de igualación utilizado para la detección del DIF con el método MH fue purificado con el procedimiento bietápico recomendado por Holland y Thayer (1988). También se utilizó este procedimiento para la purificación del criterio empleado con la RL (Navas y Gómez, 1994). Los 9 ítems con EID mostraron DIF en el primer paso durante la purificación de los criterios de igualación para los dos métodos estadísticos.

RESULTADOS

La presentación de los resultados se ha dividido en cuatro apartados: (1) el análisis de las distribuciones de puntuaciones totales y el análisis de ítems; (2) el estudio de la dimensionalidad del test de rendimiento; (3) el análisis del DIF en los ítems diseñados para medir el apartado objeto de una EID; y (4) el acuerdo entre el método MH y la RL.

1) Análisis de las distribuciones de puntuaciones totales y análisis de ítems.

El análisis de las distribuciones de puntuaciones totales permite comprobar la efectividad del procedimiento experimental. Las puntuaciones totales de las personas son el número de ítems que han contestado correctamente. La Tabla 2 muestra las medias y las desviaciones típicas por grupos en tres conjuntos de ítems.

Tabla 2. Distribuciones de puntuaciones totales.

Grupos		Items EID			Items no EID			Test completo		
	N	n	Media	DT	n	Media	DT	n	Media	DT
GR	173	9	6.70	1.87	41	26.99	4.68	50	33.69	5.82
GF	151		1.97	1.71		25.27	5.59		27.25	6.35

Los resultados fueron los esperados. La media del número de aciertos en los ítems con EID fue significativamente más alta en el GR ($t = 23.59$; $p < .001$). El GR tuvo también una media más alta en el test completo ($t = 9.53$; $p < .001$), y en el resto de los ítems ($t = 3.00$; $p = .003$), aunque en este último caso la diferencia es ligeramente superior a un punto.

Los valores del índice "p" conjunto para los dos grupos muestran que los ítems con EID son de dificultad media (están en el intervalo 0.37-0.65). Los valores del índice "p" para cada grupo revelan que los ítems con EID son más fáciles para el GR que para el GF. Las diferencias en las proporciones de acierto están en un intervalo entre 0.32 y 0.77 con un valor medio de 0.52, siendo todas significativas.

A su vez, todos estos ítems tuvieron niveles de discriminación adecuados. La media de los valores de la correlación biserial fue de 0.60.

Estos análisis prueban que la manipulación instruccional provoca las diferencias esperables en la dificultad de los ítems con EID y en las distribuciones de puntuaciones totales.

2) Dimensionalidad del test de rendimiento.

Los análisis pretendían examinar la dimensionalidad de las respuestas a todos los ítems del test y, en particular, al subconjunto de los ítems con EID.

La dimensionalidad del test en su conjunto fue analizada primero con un análisis factorial de ejes principales a partir de la matriz de correlaciones tetracórica entre los ítems. La magnitud del primer autovalor fue 8.96 y la del segundo 4.53. Aunque la diferencia es prácticamente el doble, la magnitud del segundo autovalor incita a pensar en una posible fuente de multidimensionalidad en este conjunto de datos.

La Tabla 3 muestra los resultados obtenidos con el procedimiento DIMTEST para tres conjuntos de respuestas.

Tabla 3. Dimensionalidad de los ítems.

Conjuntos de ítems	T - conservador		T' - más potente	
	T	p - valor	T'	p - valor
Todos los ítems (1)	7.9658	.0000	8.3184	.0000
Ítems con EID (2)	7.7215	.0000	8.1594	.0000
Resto de los ítems (1- 2)	-0.8595	.8049	-1.1131	.8671

Primero, se investigó la unidimensionalidad esencial del test de rendimiento. DIMTEST eligió de forma automática los ítems para el subtest de evaluación (AT1). Los valores del estadístico "T" permiten rechazar la hipótesis de que se cumpla el supuesto de unidimensionalidad esencial. A continuación, se investigó la dimensionalidad de las respuestas a los ítems con EID. La opción de DIMTEST que permite al usuario elegir los ítems para AT1 fue utilizada para formar el subtest con los ítems con EID. La Tabla 3 muestra que las respuestas a los ítems con EID no cumplen el supuesto de unidimensionalidad esencial.

Por último, se analizó el subconjunto de respuestas al resto de los ítems. La Tabla 3 indica que este subconjunto cumple el supuesto de unidimensionalidad esencial.

Los análisis de la dimensionalidad proporcionan dos argumentos relevantes para el objetivo de la investigación: (1) apuntan a la multidimensionalidad de los ítems con EID como posible explicación de su funcionamiento diferencial; y (2) refuerzan la eliminación de estos ítems de los criterios de igualación en la posterior detección del DIF.

3) Estudio de los ítems diseñados para mostrar una ejecución diferencial.

Los análisis para detectar el posible DIF de los ítems con EID fueron realizados con los procedimientos MH y RL.

Método Mantel-Haenszel

Se utilizó la purificación bietápica del criterio de igualación para la detección del DIF uniforme. El primer paso de la purificación detectó DIF en los 9 ítems con EID y en otros 5 ítems. El análisis de contenido de esos 5 ítems no aportó ninguna interpretación coherente para su funcionamiento diferencial. El criterio de igualación quedó formado por 36 ítems.

La Tabla 4 muestra los resultados de la aplicación del método MH tradicional a los 9 ítems diseñados para mostrar una ejecución diferencial. Además de los valores del estadístico MH - χ^2 y el nivel de significación, la tabla presenta los valores del índice DELTA-MH y su error de estimación.

Tabla 4. Estadísticos MH de los ítems con EID.

Nº ítem	MH-χ^2	p - valor	DELTA-MH	Error DELTA-MH
21	46.4232	.0000	-4.1602	0.6268

26	95.2742	.0000	-5.8506	0.6571
27	136.5063	.0000	-7.9517	0.7930
29	34.0737	.0000	-3.4332	0.5837
35	25.7187	.0000	-3.1102	0.5224
37	106.7768	.0000	-7.0216	0.7707
40	182.8965	.0000	-9.7891	0.9221
46	103.9945	.0000	-6.5498	0.7119
48	102.2088	.0000	-5.8724	0.7203

Nota: Los resultados del DIF son significativos con p<.001

La medida más fiable al haber utilizado una estrategia de igualación gruesa es el valor del estadístico, debiendo ser interpretados los valores del índice DELTA-MH como indicadores aproximados de la dirección y magnitud del DIF (Donoghue y Allen, 1993).

La Tabla 4 muestra que los 9 ítems diseñados para mostrar una ejecución diferencial presentaron un DIF significativo. El signo negativo de los valores del índice DELTA-MH para los 9 ítems, indica que el DIF favorece al GR.

Además, se utilizó el procedimiento de partición de la muestra sugerido por Mazor, Clauser y Hambleton (1994) para la detección de un posible DIF no uniforme. La disminución en el tamaño de la muestra obligó a formar 4 niveles en el criterio de igualación en lugar de los 5 utilizados para el método MH tradicional.

La Tabla 5 presenta los indicadores de DIF para las dos mitades en que se divide la muestra.

Tabla 5. Estadísticos MH no uniforme de los ítems con EID.

Item	MITAD INFERIOR			MITAD SUPERIOR		
	MH-χ^2	p-valor	DELTA	MH-χ^2	p-valor	DELTA
21	5.7285	.0167	-2.2632	36.9209	.0000	-5.1669
26	17.5348	.0000	-3.7884	69.2069	.0000	-7.6643
27	48.4365	.0000	-7.8160	76.6105	.0000	-8.2283
29	15.8981	.0000	-3.6417	12.8084	.0000	-3.1087
35	3.7914	.0515	-1.9751	18.9993	.0000	-3.9964
37	40.7001	.0000	-5.7463	48.3084	.0000	-8.3893
40	79.2972	.0000	-12.4799	82.1903	.0000	-9.8277
46	40.6357	.0000	-6.1935	43.8494	.0000	-6.8610
48	41.3207	.0000	-6.1935	37.2697	.0000	-4.7149

Nota: Los resultados del DIF son significativos con p<.001

Los indicadores de DIF recogidos en la Tabla 5 señalan que el ítem 21 muestra un DIF no uniforme. Se decidió representar las diferencias entre las proporciones de aciertos de los grupos en el ítem 21 y en otro ítem con DIF uniforme.

La Figura 1 representa las diferencias entre las proporciones de aciertos del GR y del GF en el ítem 21 y el ítem 46 (DIF uniforme) a lo largo de los 5 niveles del criterio de igualación. La muestra que las diferencias entre las proporciones de acierto en el ítem 21 aumentan conforme los niveles del criterio de igualación agrupan puntuaciones totales más altas, mientras que en el ítem 46 las diferencias se mantienen más constantes.

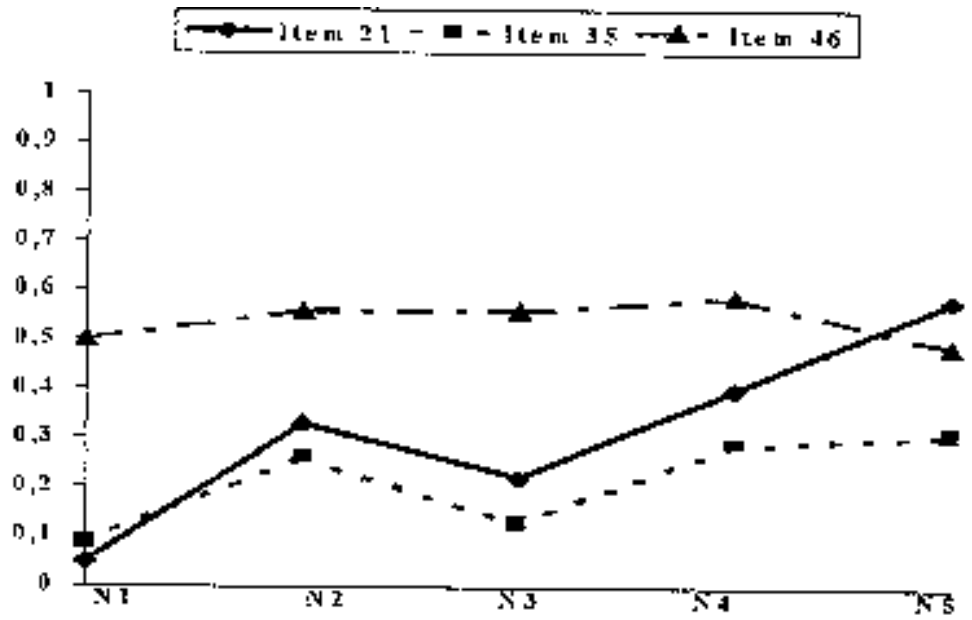

Figura 1. Diferencias entre las proporciones de aciertos del GR y del GF.

Regresión logística

El primer paso del procedimiento bietápico de purificación del criterio de igualación detectó DIF en los 9 ítems con EID y en otros 3 ítems. Tampoco en este caso se pudo elaborar una interpretación coherente para el DIF de estos tres ítems. El criterio de igualación quedó formado por 38 ítems.

La Tabla 6 muestra los valores de los estadísticos de Wald para los parámetros τ_2 y τ_3 y sus niveles de significación.

Los valores del estadístico de Wald muestran que los 9 ítems con EID presentaron un DIF significativo. Además, los ítems 21, 26 y 35 presentaron un DIF no uniforme.

El DIF de los 9 ítems diseñados para mostrar una ejecución diferencial supone una confirmación general del objetivo del estudio.

Tabla 6. Estadísticos de Wald de los ítems elaborados para mostrar DIF.

Nº ítem	τ_2	p – valor	τ_3	p - valor
21			44.0575	.0000
26			72.2888	.0000
27	99.7666	.0000		
29	33.9055	.0000		
35			25.4857	.0000
37	83.7323	.0000		
40	116.2577	.0000		
46	86.0752	.0000		
48	78.5958	.0000		

Nota: Se incluye el valor de los estadísticos para los parámetros significativos.

4) Acuerdo entre el método MH y la RL.

Tras la detección del DIF se investigó el acuerdo entre los resultados del método MH y la RL. La Tabla 7 presenta un resumen de los resultados obtenidos con los dos procedimientos.

Tabla 7. Acuerdo entre el método MH y la RL.

	MH		RL	
Items	DIF uniforme	DIF no uniforme	DIF uniforme	DIF no uniforme
Con EID	26, 27, 37, 40, 46, 48	21, 35	27, 29, 37, 40, 46, 48	21, 26, 35
Sin EID	1, 2, 8, 50	40	22, 44	17

La Tabla 7 muestra el acuerdo entre los dos procedimientos en la detección del DIF para los ítems con EID. No coinciden en la detección del DIF no uniforme en los 9 items con EID, ni con respecto al DIF no previsto.

DISCUSIÓN

El estudio se realizó para examinar el efecto de las diferencias instruccionales sobre el DIF. Se utilizó un procedimiento experimental para inducir DIF manipulando diferencialmente la instrucción recibida por los dos grupos de comparación (GR - GF). Además, se investigó el grado de acuerdo en el método MH y la RL para la detección del DIF provocado experimentalmente. Los resultados de los análisis muestran que: (1) las respuestas a los 9 ítems objeto de una estrategia instruccional diferencial no cumplen el supuesto de "unidimensionalidad esencial", y (2) que estos 9 ítems tienen un funcionamiento diferencial a favor del GR. La coincidencia entre los resultados obtenidos con los dos métodos estadísticos para detectar DIF aumenta la confianza en los resultados.

La representación gráfica de las diferencias entre las proporciones de acierto en los ítems que muestran un DIF no uniforme (21 y 35), frente al que muestra un DIF uniforme (46), sugiere que son las personas de menos habilidad las que más se benefician de la estrategia instruccional dirigida a la adquisición de un modelo mental.

El análisis subjetivo de los dos ítems que muestran un DIF no uniforme revela una semejanza en su contenido. Los tres requieren sólo recordar el tipo de interpretación de los componentes estimados de la varianza resaltada por el diagrama. No obstante, el alcance de los resultados se debe limitar a las características de un estudio exploratorio. El principal interrogante a responder es la naturaleza de las "habilidades ruido" que provocan la manipulación instruccional: familiaridad con la tarea, experiencia en responder a los ítems, facilidad para la interpretación numérica, etc.

Los dos métodos estadísticos coinciden en señalar el DIF en los 9 ítems con EID, y disienten en la identificación del DIF no uniforme en uno de los 9 ítems con EID (ítem 35). Tampoco coinciden en identificar los ítems con un DIF no previsto. El acuerdo avala la idoneidad de los métodos para identificar el DIF en un contexto experimental, pudiendose atribuir las diferencias a las peculiaridades de la situación experimental: reducido tamaño muestral, falta de potencia del método MH modificado, fiabilidad del criterio de igualación, etc.

La consideración del DIF detectado como evidencia de sesgo requiere una delimitación precisa de tres conceptos: multidimensionalidad, funcionamiento diferencial y sesgo en el ítem. Camilli y Shepard (1994) aportan los argumentos necesarios. Los índices estadísticos de FDI detectan la multidimensionalidad en las respuestas a los ítems, mientras que la consideración de que el FD detectado es una evidencia de sesgo implica: (1) identificar las causas de la ejecución diferencial en el ítem y, (2) considerar que las habilidades añadidas no son relevantes para el objetivo del test. Tal vez la utilización normativa de las puntuaciones totales para asignar calificaciones podría hacer pensar en un caso de sesgo por las diferencias instruccionales.

Esta delimitación expande el campo de aplicación de los métodos estadísticos para analizar el FDI. Además de su aplicación rutinaria dentro del análisis numérico de ítems, la utilización conjunta de métodos y análisis subjetivos puede ayudar a describir el proceso de respuesta a los ítems de rendimiento.

ABSTRACT

During the latest two decades, research on test bias has been concerned in the development of statistical methods for detecting items with a possible differential functioning. The study of causes of differential item functioning (DIF) has received less attention. This paper explores the effect of people receiving different instructional strategies on DIF. An experimental design was used to induce DIF. DIF was induced by manipulating the instruction to the two groups of examinees. The study also compared the qualities of the statistical method Mantel-Haenszel's χ^2 (Holland & Thayer, 1988) and logistic regression (Swaminathan & Rogers, 1990) for detecting DIF items. The experimental procedure was intended to induce DIF in 9 items. Both statistical methods flagged the nine DIF induced items.

Key words: DIF, causes, Mantel-Haenszel, logistic regression.

REFERENCIAS

Angoff, W. H., y Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10,* 95-106.

Camilli, G. y Shepard, L. (1994). *Methods for identifiying biased test item.* Thousand Oaks. CA: Sage Publications, Inc.

Crocker, L. y Algina, J. (1986). *Introduction to Clasical and Modern Test Theory*. Rinehart and Winston, New York.

Donoghue, J. R. y Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18,* 131-154.

Fidalgo, A. M. (1966). Funcionamiento diferencial de los items. En J. Muñiz (cord.), *Psicometría* (pp. 371-457). Madrid: Editorial Universitas, S. A.

Hambleton, R. K., Clauser, B. E., Mazor, K. M. y Jones, R. W. (1993). Advances in the detection of differential functioning test items. *European Journal of Psychological Assessment, 9,* 1-18.

Hattie, J., Krakowski, K., Rogers, H. J. y Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20,* 1-14.

Holland, P. W. y Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. En H. Wainer y H. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Kok, F. G., Mellenbergh, G. J. y Van der Flier, H. (1985). Detecting experimentally induced item bias using the Iterative Logit method. *Journal of Educational Measurement, 22,* 295-303.

Linn, R. L. y Harnish, D. L. (1981). Interaction between Item content and group menbership on achievement test items. *Journal of Educational Measurement, 18,* 109-118.

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

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127-163.

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.

Muthén, B. O. (1988). Some uses of structural causation modeling in validity studies extending IRT to external variables. En H. Wainer y H. Braum (Eds.), *Test Validity* (pp. 213-238). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Nandakumar, R. (1991). Traditional dimensionality vs. essential dimensionality. *Journal of Educational Measurement, 28,* 99-117.

Nandakumar, R. (1994). Assessing dimensionaliy of a set of item responses. Comparison of different approach, *Journal of Educational Measurement, 31,* 17-35.

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

O'Neill, K. A. y McPeek, W. M. (1993). Item and test characteristics that are associated with Differential Item Functioning. En P.W. Holland y H. Wainer (Eds.), *Differential ítem functioning* (pp. 255-277). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Padilla, J.L., Pérez, C., González, A. (1998). La explicación del sesgo en los ítems. *Psicothema, 2,* 481-490.

Padilla, J.L., Pérez, C., González, A. (1998). La dimensionalidad del test y las diferencias instruccionales. *Psicothema* (en prensa).

Raju, N.S., Bode, R.K. y Larsen, V.S. (1989). An empirical assesment of the Mantel-Haenszel statistic for sStudying differential item performance. *Applied Psychological Measurement, 2,* 1-13.

Scheuneman, J. D. (1982). A posteriori analyses of biased items. En R.A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 64-96). Baltimore. Maryland: The Johns Hopkins University Press.

Scheuneman, J. D. (1987). An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement, 24,* 97-118.

Scheuneman, J. D. y Gerritz, K. (1990). Using differential item functioning procedures to explore sources of items difficulty and group performance characteristics. *Journal of Educational Measurement, 27,* 109-131.

Schmitt, A.P. y Dorans, N.J. (1990). Differential item functioning for minority examinees on the SAT. *Jounal of Educational Measurement, 27,* 67-81.

Schmitt, A. P., Holland, P. W. y Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. En P.W. Holland y H. Wainer (Eds), *Differential Item Functioning* (pp. 281-313). Hillsdale, New Jersey: Lawrence Erlbaum Associtates, Publishers.

Skagg, G. y Lissitz, R. W. (1992). The consistency of detecting item bias across different test administrations: Implications of another failure. *Journal of Educational Measurement, 29,* 227-242.

Schmeiser, C. B. (1982). Use of experimental design in statistical item bias studies. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 64-96), Baltimore, Maryland: The Johns Hopkins University Press.

SPSS, Inc. (1993). *SPSS for windows* (Version 6.0) [Computer software]. Chicago: Author.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589-617.

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

Tatsuoka, K. K., Linn, R. L., Tatsuoka, M. M. y Yamamoto, K. (1988). Differential item functioning resulting from the use of different solution strategies. *Journal of Educational Measurement, 25,* 301-319.

Zorroza, J. y Sánchez-Cánovas, J. (1995). Los componentes cognitivos de la capacidad matemática: Representación mental, esquemas, estrategias y algoritmos. *Psicológica,* 16, 305-320.

(Revisión aceptada: 7/10/98)