Psicothema, 1998. Vol. 10, nº 2, pp. 481-490
ISSN 0214 - 9915 CODEN PSOTEG

LA EXPLICACIÓN DEL SESGO EN LOS ÍTEMS DE RENDIMIENTO

José Luis Padilla García, Cristino Pérez Meléndez y Andrés González Gómez
Universidad de Granada

La identificación de los factores responsables del sesgo en los items no ha alcanzado las expectativas iniciales. La investigación para determinar las causas ha seguido dos líneas interrelacionadas: las características de los items o las características diferenciales de las personas. Presentamos un estudio empírico que explora el efecto de recibir diferentes experiencias instruccionales sobre el sesgo en los items de rendimiento. El sesgo es intencionadamente inducido manipulando la estrategia instruccional. La adscripción de las personas a los dos grupos de comparación depende de la estrategia instruccional recibida. El procedimiento estadístico χ^2 de Mantel Haenszel detecta el posible funcionamiento diferencial de 10 items especialmente diseñados para mostrar sesgo. Las implicaciones de los resultados son discutidas desde la perspectiva de la teoría de la validez.

The influence of instructional experience on achievement item bias. Identification of factors responsible for item bias has not reached initial expectations. Research to determine its causes has followed two main interrelated lines: items characteristic or differential characteristics of people. We present an empirical study that explores the effect of receiving different instructional experiences on achievement item bias. Manipulating instructional strategy intentionally induces bias. The assignment of people to the comparison groups depends on the instructional strategy received. Mantel-Haenszel's statistical procedure χ^2 detects a possible differential functioning of 10 items especially designed to display bias. Implications of results are addressed from the validity theory perspective.

Desde los inicios del estudio del sesgo en los ítems, ha estado presente la preocupación por identificar los factores responsables del mismo (Angoff y Ford, 1973). Tras aproximadamente dos décadas de estudio, la evaluación mayoritaria coincide en que los logros conseguidos son decepcionantes: el trabajo metodológico se ha convertido en un fin en sí mismo (Scheuneman, 1987); la predicción del posible sesgo de un item es muy difícil (Skagg y Lissitz, 1992); los factores responsables conocidos son muy pocos y la comprensión de cómo actúan limitada (Schmitt, Holland y Dorans, 1993), etc. La justificación principal para esta ausencia de logros significativos puede estar en una concepción teórica inicial sobre las posibles causas demasiado simple: la hipótesis de la carga cultural.

Correspondencia: José Luis Padilla García
Facultad de Psicología
Universidad de Granada
18071 Granada (Spain)
E-mail: jpadilla@platon.ugr.es

Esta situación general no debe ocultar los esfuerzos realizados para identificar los factores responsables del sesgo en los ítems de rendimiento. Scheuneman (1982) agrupa los trabajos en dos categorías dependiendo de las posibles fuentes de sesgo analizadas: (a) defectos en los ítems a los que los miembros de grupos diferentes son diferencialmente sensibles; y (b) diferencias genuinas entre los grupos que pueden o no reflejar diferencias válidas en la habilidad medida.

El objetivo del primer grupo de trabajos es identificar características irrelevantes de los ítems que den lugar a estimaciones equivocadas de la habilidad de las personas. Este es el objetivo general compartido de los trabajos de Angoff y Ford (1973), Linn y Harnish (1981), O'Neill y McPeek (1993), Scheuneman y Gerriz (1990) y Schmitt y Dorans (1990). La mayoría de estos trabajos concluyen que las diferencias en las experiencias instruccionales de los miembros de cada grupo son las responsables del funcionamiento diferencial de los ítems (FDI).

El papel de las diferencias instruccionales es estudiado directamente por Miller y Linn (1988). Investigaron el grado en el que las funciones características de los ítems eran invariables entre grupos con diferentes experiencias instruccionales. Los resultados mostraron que entre el 20% y el 30% de los ítems no lograban la invarianza. Kok, Mellenbergh y van der Flier (1985) realizaron una investigación especialmente interesante por mostrar la posibilidad de inducir sesgo de forma controlada, manipulando los procesos instruccionales que seguían dos grupos de personas. Tatsuoka, Linn, Tatsuoka y Yamamoto (1988) compararon la ejecución en los ítems de un test de sustracción de fracciones de dos grupos de niños, formados a partir de los procesos cognitivos reales –método de cálculo– que empleaban para resolver las tareas. Los ítems que requerían un método particular de cálculo favorecían a los alumnos que empleaban dicho método. Muthén (1989) formuló un modelo de TRI que incorpora información sobre la instrucción. Propone la idea de que la instrucción mejora la «habilidad objetivo» y, posiblemente, las destrezas específicas –«habilidades ruido»– que pueden favorecer la elección de la respuesta correcta a algunos ítems.

Los resultados de esta línea de investigación son aún insuficientes para sostener una teoría global, si fuera posible, sobre la explicación del FDI. Nos atrevemos a apuntar algunas razones de la falta de resultados consolidados. Por ejemplo: (a) la mayoría no son estudios diseñados intencionadamente para examinar las causas del sesgo, suele tratarse de análisis «a posteriori»; (b) los tests analizados son tests comerciales que han pasado numerosos controles, por lo que rara vez aparecen ítems con un FD significativo; y (c) predomina el análisis subjetivo de los ítems, por lo que falta la manipulación controlada de las posibles fuentes de sesgo.

Este estudio pretende explorar el efecto de las diferencias en las experiencias instruccionales sobre el sesgo en los ítems de rendimiento. La lógica del estudio, inspirada en la «teoría multidimensional del sesgo» (Fidalgo, 1996), es que las diferencias en las experiencias instruccionales pueden provocar diferencias en la «habilidad objetivo» (rendimiento en el área de contenido), y en las «habilidades ruido» (destrezas específicas para responder al item). La presencia de habilidades ruido provocará que las respuestas a los ítems no sean unidimensionales, y la aparición de un FD en los ítems afectados por dichas habilidades.

El objetivo del estudio se puede concretar en dos proposiciones interrelacionadas: (a) obtener evidencias para determinar si recibir o no una estrategia instruccional puede ser una factor explicativo del FD de los ítems de un test de rendimiento; y (b) examinar si el posible FD de los ítems de ren-

dimiento puede proceder de la multidimensionalidad que las diferentes experiencias instruccionales originan en las respuestas a los ítems.

Método

Sujetos y diseño

Participaron 336 personas de ambos sexos. Todas cursaban la asignatura de Psicometría dentro del tercer curso de la Licenciatura de Psicología. El área de contenido utilizada para la experiencia instruccional fue el tema «Análisis numérico de ítems». Las personas fueron asignadas al azar a dos grupos: 165 sujetos al grupo de referencia (GR) y 171 al grupo focal (GF).

Experiencia instruccional

La oportunidad de aprender (ODA) ha sido la variable instruccional tradicionalmente empleada por los psicómetras para obtener información sobre las experiencias instruccionales de las personas. La ODA es una variable dicotómica cuyos valores dependen de que las personas hayan tenido o no, oportunidad de aprender el contenido examinado en los ítems (Muthén, 1989).

Sin embargo, según algunos autores (i.e. Miller y Linn, 1988) la ODA podría ocultar la dinámica de la enseñanza en las aulas. Por esta razón, decidimos sustituir la ODA por otra variable instruccional: haber recibido o no una enseñanza dirigida a favorecer la adquisición de un modelo mental sobre un contenido instruccional. Numerosos estudios (i.e. Gagné, 1987) muestran que, durante el aprendizaje, las personas elaboran representaciones –«modelos mentales»– que dirigen su ejecución en tareas de evaluación. Los modelos mentales incluyen información sobre los requisitos de la tarea y cómo realizarla (Gagné y Glaser, 1987). La enseñanza dirigida a la adquisición de un modelo mental utiliza diagramas, ejemplos y no-ejemplos adecuados para las tareas que deberán resolver los alumnos.

En definitiva, la manipulación instruccional consiste en realizar una enseñanza dirigida a favorecer la adquisición de un modelo mental para mejorar la ejecución en un conjunto específico de ítems, mediante la adquisición de supuestas «habilidades ruido», que pueden facilitar la elección de la respuesta correcta al item.

Materiales

Los materiales son un cuestionario, las unidades de tratamiento y un test de rendimiento. Los instrumentos de medida fueron elaborados por los autores, siguiendo los pasos establecidos en la literatura (Osterlind, 1989).

1) Cuestionario

El cuestionario proporcionó información sobre variables que podían influir en la ejecución de las personas en los ítems del test de rendimiento: datos demográficos, actitud hacia la estadística y conocimientos previos de estadística descriptiva.

2) Unidades de tratamiento

Las unidades de tratamiento son dos informes escritos que presentan básicamente la misma información sobre el tema «Análisis numérico de ítems». Ambos informes, recogen los contenidos del tema tal y como aparecen en los manuales de Psicometría más conocidos (e.g. Crocker y Algina, 1986; Magnusson, 1968; Osterlind, 1989)

Los informes que recibían los dos grupos diferían en el modo de presentación. Estas diferencias se limitan a dos apartados del tema: «La utilización del índice ‘p’ en el análisis de ítems» y «La comparación de los índices ‘p’ obtenidos por subgrupos de sujetos con altos y bajos niveles de habilidad», ya que son estos los apartados sobre los que se deseaba realizar una estrategia instruccional

diferencial (EID). En concreto, en el informe que se entregaba a los sujetos del GR, la presentación de estos apartados, además de hacerse de forma descriptiva se acompañaba de diagramas (1 diagrama principal y cuatro diagramas parciales), ejemplos y no-ejemplos. El diagrama principal representa un modelo que describe la utilización del índice 'p' para analizar la calidad de un item. Los ejemplos y no-ejemplos interpretan los resultados del índice 'p' para las alternativas de respuesta al item. La interpretación analiza la elección de la respuesta correcta y los distractores. En el GF la presentación de los contenidos era fundamentalmente descriptiva. La secuencia de presentación de los contenidos fue la misma en los dos informes.

La Tabla 1 presenta el esquema del proceso instruccional que reciben los sujetos en el apartado del contenido objeto de una EID.

Tabla 1
Proceso Instruccional

Grupos	Estrategia instruccional	Modo de presentación	Número de ítems
Grupo de Referencia	Adquirir modelo mental relevante	5 diagramas + 9 ejemplos + 4 no-ejemplos	10 items
Grupo Focal	Presentación descriptiva	4 ejemplos	

La elaboración de los informes se hizo de acuerdo con la aproximación al diseño instruccional basada en los trabajos sobre el tema (e.g. Gagné, Briggs y Wager, 1988; Merrill, Tennyson y Posey, 1992).

3) TEST DE RENDIMIENTO

El test de rendimiento fue elaborado para medir la ejecución de las personas en el tema del «Análisis numérico de ítems». Estaba formado por 37 ítems de elección múltiple con tres alternativas de respuesta. El test contenía 10 ítems diseñados para medir los apartados del contenido objeto de una estrategia instruccional diferencial. Los ítems demandan la interpretación de los resultados del índice 'p' de la forma presentada en los ejemplos y no-ejemplos de las unidades de tratamiento. El número de ítems «potencialmente sesgado» se ajustaba a los criterios expuestos por diversos autores para evitar la aparición de un «sesgo penetrante» (Oshima y Miller, 1992).

Procedimiento

El estudio del contenido de los informes, la administración del cuestionario y el test de rendimiento, se realizaron en sesiones de grupo. Las personas no conocían los objetivos del estudio.

Tras repartir los informes, se pedía a las personas para que estudiaran el contenido como «si se estuvieran preparando para un examen...». También se les informaba que después de estudiar el material iban a responder a un cuestionario y a un test sobre los contenidos estudiados. Después de estudiar el material respondían al test.

Resultados

El análisis de las variables que podían influir en la ejecución de las personas en el test de rendimiento mostró que: (a) el 86,2% de las personas era la primera vez que cursaban la asignatura de Psicometría; (b) no había diferencias significativas entre los grupos con respecto a la actitud hacia la estadística ($t= -1.27$; $p= 0.20$), ni en cuanto a los conocimientos previos de estadística descriptiva ($t= 0.36$; $p= 0.76$).

El bloque principal de resultados de los análisis lo hemos dividido en dos apartados: (a) el estudio de la dimensionalidad del test de rendimiento; y (b) la aplicación del procedimiento χ^2 de Mantel-Haenszel, para el análisis de los ítems diseñados para mostrar sesgo.

1) DIMENSIONALIDAD DEL TEST DE RENDIMIENTO

El análisis de la dimensionalidad del test de rendimiento tiene un doble objetivo. Primero, determinar si el conjunto de respuestas al test de rendimiento es unidimensional. Segundo, determinar si las respuestas a los ítems diseñados para mostrar sesgo forman un conjunto multidimensional.

El análisis de la unidimensionalidad se realizó mediante la aplicación del estadístico «T» de Stout (e.g. Stout, 1987), por su relevancia para los objetivos de la investigación y para evitar los problemas del análisis factorial al estudiar la dimensionalidad de un test de rendimiento (e.g. Cuesta, 1993). El estadístico «T» de Stout prueba la hipótesis de «unidimensionalidad esencial» en el conjunto de respuestas a los ítems del test. El cálculo del valor del estadístico «T» se realizó con el programa DIMTEST (Stout, l990).

El estadístico asume el principio fundamental de que la independencia local se debería cumplir aproximadamente cuando la muestra procede de un subpoblación de personas con aproximadamente el mismo nivel de habilidad. Así, un test es esencialmente unidimensional si el promedio de las covarianzas condicionadas entre todas las parejas de ítems es pequeño.

El procedimiento consiste en formar un conjunto de ítems unidimensionales, llamado *subtest de evaluación*, mediante un análisis subjetivo del contenido de los ítems o mediante un análisis factorial exploratorio. El resto de los ítems forman el *subtest de agrupación*. Después, las personas son asignadas a diferentes grupos por sus puntuaciones en los ítems del subtest de agrupación. Si el conjunto total de ítems es undimensional, ambos subtest serán unidimensionales, pero si no es así, el subtest de agrupación contendrá muchos ítems que «cargarán» en al menos otra dimensión no medida por el subtest de evaluación (Stout, 1987).

El procedimiento DIMTEST ha sido analizado en numerosos estudios a los que se puede recurrir para conocer los detalles de su aplicación (e.g. Nandakumar, 1993, 1994; Hattie, Krakowski, Rogers y Swaminathan, 1996).

La Tabla 2 muestra los resultados de la aplicación del estadístico «T» de Stout al conjunto de todos los ítems del test. El programa asignó automáticamente los ítems que formaron el subtest de evaluación (AT1).

Tabla 2
Estadístico «T» para el test completo

T-Conservador				T'-Más potente			
TL	TB	T	p-valor	TL	TB	T'	p-valor
3.5629	1.7628	1.2728	0.1015	4.2550	2.1499	1.4885	0.0683

El valor del estadístico de Stout, tanto en su versión conservadora (T= 1.2728, p= .1015) como en la más potente (T'= 1.4885, p= .06831), no permite rechazar la hipótesis nula de que en el conjunto de datos se cumple el supuesto de unidimensionalidad esencial.

Para determinar si la respuesta de los sujetos a ítems objeto de EID forman un conjunto multidimensional, aplicamos también el estadístico «T» de Stout. En esta ocasión, se utilizó la opción de asignación «basada en la opinión de expertos» del programa DIMTEST para formar AT1 (i.e. los ítems diseñados para mostrar un FD).

La Tabla 3 muestra los resultados de la aplicación del estadístico «T» de Stout.

Tabla 3
Estadístico «T» para las respuestas a los items objeto de EID

T-Conservador				T'-Más potente			
TL	TB	T	p-valor	TL	TB	T'	p-valor
6.8536	1.7718	.35935	.0001	7.6476	2.1517	3.8862	.00005

Los valores del estadístico «T» de Stout, tanto en su versión conservadora (T= 3.5935, p= .000162), como en la más potente (T= 3.8862, p= .000051), permiten rechazar la hipótesis nula de que en el conjunto de datos formado por los ítems objeto de EID se cumple la unidimensionalidad esencial. También puede destacarse el incremento en los valores del estadístico TL (i.e. medida de la multidimensionalidad presente en las respuestas a los ítems del subtest de evaluación uno), para los ítems objeto de EID (TL= 6.8536), en comparación con el resto de los ítems del test (TL= 3.5629).

Este resultado, junto con el anterior, puede interpretarse así: a pesar de que la unidimensionalidad esencial se cumple para el conjunto de datos del test, también en este conjunto existen dimensiones menores que, en el caso de los ítems objeto de EID, pueden dar lugar a su posible FD.

3) Estudio de los ítems diseñados para mostrar sesgo

El análisis de los ítems diseñados para mostrar sesgo comenzó con dos controles estadísticos relevantes para la aplicación del procedimiento χ^2 de Mantel-Haenszel (MH): el análisis numérico de los ítems y de las distribuciones de puntuaciones totales en el test de rendimiento.

Los valores del índice 'p' en los dos grupos de todos los ítems objeto de una EID indican que estos ítems, salvo el item 20, son más fáciles para el GR que para el GF. Las diferencias en los valores del índice 'p' son considerables (superiores a .30) para los ítems 1, 2, 12, 24 y 28; y apreciables (superiores a .10) para los ítems 11 y 33. Estas diferencias en los valores del índice 'p' revelan la efectividad de la manipulación instruccional, y se producen en la dirección prevista.

Podemos resaltar también, que todos los ítems objeto de EID, salvo nuevamente el item 20, tienen niveles de discriminación adecuados (i.e. por encima o próximos a r_{bis}= .50); dichos niveles son una garantía más de la identidad entre lo que mide el item y la distribución de puntuaciones totales (Angoff, 1993). El análisis subjetivo del item 20 aclaró que su comportamiento anómalo podía deberse a que no era congruente con el objetivo educativo que se pretendía medir.

Por lo que respecta al análisis de las distribuciones de puntuaciones totales, realizamos dos contrastes de diferencias de medias, con y sin los ítems diseñados para mostrar sesgo, entre los dos grupos. La inclusión de los ítems diseñados para mostrar sesgo favorecía al GR de forma significativa ($t = 4.26$; $p<.001$); mientras que con su exclusión las distribuciones de puntuaciones totales eran muy semejantes ($t = 1.29$; $p = .1995$).

Estos resultados y la reflexión teórica determinaron el protocolo de aplicación del procedimiento:

1) La estrategia de igualación. La estrategia típica es la utilización simple de la puntuación total en el test (para n ítems, n+1 niveles en la variable de igualación). La alternativa utilizada en este estudio es una de las estrategias de igualación gruesa: los quintiles de la distribución conjunta de puntuaciones totales. Las razones para esta decisión fueron tres: (a) conseguir mayor estabilidad en las estimaciones de las frecuencias esperadas; (b) utilizar la mayor parte de los datos disponibles, reduciendo el número de filas y columnas con frecuencia cero; y (c) ser la igualación gruesa más cercana a la tradicional igualación delgada, en cuanto a la comparabilidad de las personas.

2) La inclusión de los ítems estudiados en el criterio de igualación es recomendada por la mayoría de los autores para establecer el paralelismo entre el procedimiento MH y la detección del FDI desde el modelo de

Rasch. Nosotros decidimos excluirlos para obtener un criterio de igualación lo más unidimensional posible.

Primero, aplicamos el procedimiento MH al conjunto de ítems diseñados para medir los contenidos objeto de una estrategia instruccional diferencial (EID); y después, al resto de los ítems del test.

La estrategia de igualación utilizada hace que los valores del estadístico MH-CHI^2 sean más fiables que los del estimador DELTA-MH (D-MH) (Donoghue y Allen, 1993); siendo estos últimos meros indicadores de la «dirección» y «magnitud» del FD del item estudiado.

La Tabla 4 presenta los resultados de la aplicación del procedimiento MH a los ítems diseñados para medir los contenidos objeto de EID.

Tabla 4
Aplicación del procedimiento MH a los items diseñados para mostrar sesgo

Nº del ítem	MH-CHI^2	p-valor	DELTA-MH	Error DELTA-MH	Valoración DELTA-MH
1	63.2422	.0000	-6.8289	1.0335	Grande
2	113.3036	0.0000	-8.3263	0.9575	Grande
6	11.1198	0.0091	-3.9948	1.2342	Grande
11	8.2914	0.0029	-3.6067	1.2104	Grande
12	36.0907	0.0091	-5.8881	1.1181	Grande
20	2.9689	0.0849	1.1874	0.6433	Moderado
24	68.8180	0.0000	-3.5170	0.7133	Grande
28	54.4301	0.0000	-4.3949	0.6170	Grande
31	3.9760	0.0462	-1.1103	0.5355	Moderado
33	13.4063	0.0003	-3.9395	1.1127	Grande

Los valores negativos de D-MH indican que el item favorece al GR. Según la clasificación del Educational Testing Service (ETS), valores D-MH menores de 1 en valor absoluto se considera un FDI despreciable, valores D-MH mayores de 1,5 en valor absoluto indican un FDI grande, y los ítems con valores D-MH intermedios se clasifican como un FDI moderado (Dorans y Holland, l993).

Como puede verse en la Tabla 4, 9 de los 10 ítems diseñados para medir los contenidos objeto de una EID, muestran un FD significativo. Los valores D-MH de los 9 ítems con un FD significativo son negativos, lo que indican que favorecen al grupo de referencia.

El item 20, a pesar de estar diseñado para medir un contenido objeto de una EID, no presenta un FD significativo para ninguno de los dos grupos (MH-CHI^2 = 2.9689, p = 0.0849).

Este patrón de resultados es una confirmación general de las predicciones expuestas en la introducción del estudio.

Por último, el procedimiento MH detecto un posible funcionamiento diferencial de otros 5 ítems del test de rendimiento. El análisis subjetivo de los ítems no reveló ningún patrón significativo en estos ítems.

Discusión

Este estudio se realizó para examinar el efecto de recibir diferentes estrategias instruccionales sobre el FD de los ítems de rendimiento. El análisis de las respuestas a los ítems diseñados para mostrar sesgo reveló que 1) las respuestas a los ítems forman un conjunto multidimensional; y 2) los ítems están sesgados a favor del grupo que recibe la instrucción dirigida a la adquisición de un modelo mental relevante. Dicha instrucción favorece la adquisición de «habilidades ruido» que afectan a la ejecución en el item.

El análisis de la dimensionalidad del test de rendimiento es especialmente relevante por dos razones. Primera, la unidimensionalidad del conjunto de las respuestas al test apoya el que los resultados de la aplicación del procedimiento MH no están contaminados por la hipotética multidimensionalidad del criterio de igualación. Segunda, la falta de unidimensionalidad en las res-

puestas a los ítems diseñados para mostrar sesgo proporciona evidencia sobre el posible mecanismo explicativo del sesgo: la actuación de «habilidades ruido» distribuidas de forma diferente entre los grupos de comparación.

El procedimiento MH identificó el posible funcionamiento diferencial de 9 de los 10 ítems diseñados para mostrar sesgo. Este resultado es una confirmación general de las predicciones apuntadas en la introducción del estudio.

No obstante, la falta de fiabilidad común a los procedimientos estadísticos para detectar el FDI, obliga a un análisis de la aplicación realizada en este estudio del procedimiento MH, centrada en la idoneidad del criterio de igualación.

La aplicación que hemos realizado del procedimiento MH ha pretendido utilizar un criterio de igualación tan fiable y unidimensional como fuera posible. De ahí, la utilización de una estrategia de igualación gruesa y la exclusión de los ítems diseñados para mostrar sesgo del criterio de igualación. Hay diversos estudios en los que la estrategia de igualación gruesa aporta estimaciones precisas de los índices de FDI (Raju, Bode y Larsen 1989); y los mejores resultados cuando la medida de FDI es el estadístico MH-CHI^2 (Doneghue y Allen, 1993). Además, Hambleton, Clauser, Mazor y Jones (1993) han encontrado que si las distribuciones de habilidad son semejantes, las diferencias entre los resultados con las distintas estrategias de igualación son mínimas.

La exclusión del item estudiado del criterio de igualación pretende que este sea lo más unidimensional posible. Hambleton et al. (1993) han encontrado efectos perversos despreciables en tests suficientemente largos (a partir de 20 ítems), y cuando las distribuciones de habilidad son semejantes.

Sin duda, el tamaño de los grupos es reducido, pero pensamos que los posibles problemas ocasionados son el precio justo por el control de las experiencias instruccionales en un estudio exploratorio.

La interpretación general de los resultados del estudio se debe hacer en el marco de la Teoría de la Validez. Camilli (1992) piensa que los índices de FDI son medidas de la multidimensionalidad presente en las respuestas a los ítems. El análisis de la validez de constructo debe determinar si la multidimensionalidad encontrada por los procedimientos estadísticos es una parte legítima del constructo o es una evidencia de sesgo. Creemos que nuestro estudio se presta a una reflexión de este tipo. Los resultados encontrados son una evidencia de sesgo para la utilización tradicional de los tests de rendimiento: la interpretación normativa de la ejecución en el test. Sin embargo, podría ser una «multidimensionalidad relevante» para otras interpretaciones dictadas por el contexto particular de utilización del test, por ejemplo: una interpretación relativa a objetivos educativos concretos.

La distinción entre «multidimensionalidad relevante» y «sesgo» (Camilli y Shepard, 1994) supera polémicas históricas sobre la utilización de los métodos estadísticos. Así, la decisión de si el FD detectado es o no sesgo, dependerá del objetivo para el que se utilicen las puntuaciones, y no simplemente, de que se conozcan o no las causas de la ejecución diferencial de las personas. Además, abre nuevas perspectivas a la utilización de los procedimientos estadísticos para detectar FD. Por ejemplo, para estudiar los cambios en el significado subyacente de lo que mide un item cuando se comparan grupos de personas con diferentes historias instruccionales. Este trabajo puede ser un ejemplo de esta posibilidad.

Referencias

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. En P.W. Holland y H. Wainer (Eds.), *Differential item functioning*, (pp. 3-23). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

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. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Journal of Educational Measurement*, 16, 129-147.

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.

Cuesta, M. (1993). *Utilización de modelos logísticos unidimensionales con datos multidimensionales.* Tesis doctoral no publicada, Universidad de Oviedo. Oviedo. España.

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.

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*, (pp. 35-66). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Fidalgo, A. M. (1996). Funcionamiento diferencial de los ítems. En J. Muñiz (cord). Psicometría. (pp. 371-457). Madrid: Editorial Universitas, S.A.

Gagné, R. M. (1987). *Instructional Technology: Foundation*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Gagné, R. M. y Glaser, R. (1987). Foundations in learning research. En R. M. Gagné (Ed.) *Instructional Technology: Foundation*. (pp. 49-84). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Gagné, R. M., Briggs, L. J. y Wager, W. W. (1988). *Principles of instructional design*. (3ed). New York: Holt, Rinehart y Winston.

Hambleton, R. K., Clauser, B. E., Mazor, K. M. y Jones, R. W. (1993). Advances in the detection of differential functioning test ítems. *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 ítems. *Journal of educational measurement*. 18, 109-118.

Magnusson, D. (1968). Teoría de los tests. Trillas. Mexico.

Merrill, M. D., Tennyson, R. D. y Posey, L. O. (1992). *Teaching concepts: An instructional design guide*. Englewood Cliffs, New Jersey: Educational Technology Publications, Inc.

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. (1989). Using item-specific instructional information in achievement modeling. *Psychometrika*. 135-396.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement*. 17, 29-38.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses. Comparasion of different approaches. *Journal of Educational Measurement*. 31, 17-35.

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 item functioning.* (pp. 255-277). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Oshima, T. C. y Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement*. 16, 237-248.

Osterlind, S. J. (1989). *Constructing test ítems*. Norwell. Massachusetts: Kluwer Academic Publishers.

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

Scheuneman, J. D. (1982). A posteriori analyses of biased ítems. 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 ítems. *Journal of Educational Measurement*. 24, 97-118.

Scheuneman, J. D. y Gerriz, K. (1990). Using differential ítems functioning procedures to explore sources of ítems 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.

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

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*. 55, 293-325.

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.

Aceptado el 29 de octubre de 1997