Elosua, P. (2006). Funcionamiento diferencial del ítem en la evaluación internacional PISA. Detección y comprensión. *RELIEVE, v. 12*, n. 2, p. 247-259. http://www.uv.es/RELIEVE/v12n2/RELIEVEv12n2_4.htm

RELIEVE

Revista ELectrónica de Investigación y EValuación Educativa

FUNCIONAMIENTO DIFERENCIAL DEL ITEM EN LA EVALUACIÓN INTERNACIONAL PISA. DETECCIÓN Y COMPRENSIÓN

[Differential Item Functioning in the PISA Project: Detection and Understanding]

por

Article record
About authors
HTML format

Paula Elosua (paula.elosua@ehu.es)

Ficha del artículo
Sobre los autores
Formato HTML

Abstract

This report analyses the differential item functioning (DIF) in the *Programme for Indicators of Student Achievement* PISA2000. The items studied are coming from the Reading Comprehension Test. We analyzed the released items from this year because we wanted to join the detection of DIF and its understanding. The reference group is the sample of United Kingdom and the focal group is the Spanish sample. The procedures of detection are Mantel-Haenszel, Logistic Regression and the standardized mean difference, and their extensions for polytomous items. Two items were flagged and the post-hoc analysis didn't explain the causes of DIF entirely.

Keywords

Differential Item Functioning, PISA, Mantel-Haenszel, Logistic Regression, Polytomous DIF, Test adaptation.

Resumen

Este trabajo analiza el funcionamiento diferencial del ítem (FDI) de la prueba de comprensión lectora de la evaluación PISA2000 entre la muestras del Reino Unido y España. Se estudian los ítems liberados con el fin de aunar las fases de detección del FDI con la comprensión de sus causas. En la fase de detección se comparan los resultados de los procedimientos Mantel-Haenszel, Regresión Logística y Medias Estandarizadas en sus versiones para ítems dicotómicos y politómicos. Los resultados muestran que dos ítems presentan funcionamiento diferencial aunque el estudio post-hoc llevado a cabo sobre su contenido no ha podido precisar sus causas.

Descriptores

Funcionamiento Diferencial del ítem, PISA, Mantel-Haenszel, Regresión Logística, FDI politómico, Adaptación de tests.

1. INTRODUCCIÓN

El incremento y alcance de las evaluaciones internacionales como los proyectos OCDE/PISA (Organización para la Cooperación y el Desarrollo Económico/ *Programme for Indicators of Student Achievement)* y TIMMS (*Trends in International Mathematics and Sciency Study*), en los que participan más de 30 países, son tal vez los mejores indicadores de la necesidad de estudios que desde una doble perspectiva, metodológica y sustantiva, analicen la equivalencia entre las versiones de las pruebas a utilizar.

Dentro del proceso de estudio de la equivalencia entre tests la detección del funcionamiento diferencial del ítem (FDI) es una etapa ineludible. Un ítem presenta funciona-

miento diferencial cuando la probabilidad de respuesta correcta no depende únicamente del nivel de la persona en el rasgo intencionadamente medido por el test. En esta situación la probabilidad de respuesta correcta a un ítem condicionada sobre el nivel de habilidad (q) podría ser diferente para personas pertenecientes a diferentes grupos ($P(X_i=1|q, Grupo1)$[1] $P(X_i=1|q, Grupo2)$) , lo cual infringe el supuesto de invarianza de medida. Por ejemplo, en el ámbito de la evaluación educativa, un ítem perteneciente a una escala de aptitud numérica presentaría funcionamiento diferencial cuando la respuesta correcta dependiera del nivel de los estudiantes en la variable "aptitud numérica" y de variables extrañas a los objetivos propuestos, como por ejemplo "momento de administración de la prueba", sexo de los estudiantes, currículo del alumno o "idioma de aplicación" entre otros. Una investigación llevada a cabo por López y Elosua (2002) sobre ítems de aptitud numérica reveló la presencia de FDI en función del momento temporal de la aplicación de la prueba. La tarea del ítem consistía en el reconocimiento de la escritura de los "números romanos"; dado que el estudio del tema de los "números romanos" no es un eje de fuerza dentro del sistema educativo los alumnos con una distancia temporal menor a la presentación del ítem tuvieron probabilidades de respuesta correcta mayores que el resto. En esta situación es oportuno cuestionarse la pertinencia o relevancia de ese ítem para la consecución de los objetivos perseguidos por el test. Una segunda situación común al entorno educativo actual en el que coexisten contextos lingüísticamente bilingües vendría definida por el idioma de aplicación del ítem. El idioma de presentación podría ser una fuente potencial de sesgo (Elosua, López, Egaña, Artamendi y Yenes, 2000). Un ítem podría presentar funcionamiento diferencial cuando se administre en un idioma que no es el idioma de escolarización de los estudiantes que vayan a resolverlo. En este situación, más real que hipotética, el idioma de presentación del ítem sería una variable extraña que podría distorsionar el significado de las puntuaciones. Con que factor se relaciona ese ítem? Es aptitud numérica o conocimiento del idioma? Tanto en un supuesto como en el otro, el funcionamiento diferencial del ítem reflejaría un sesgo de medida; es decir, a un error sistemático, que altera el significado de las puntuaciones e imposibilita que puedan llevarse a cabo comparaciones de ningún tipo.

En el caso de la adaptación de tests el tema no es baladí. Admitiendo que las adaptaciones literales no garantizan en ningún caso la equivalencia psicométrica, el estudio del funcionamiento diferencial del ítem se torna en una fase ineludible en el proceso de adaptación de tests (Hambleton, 2001; Hambleton, Merenda y Spiellberg, 2005) reconocido por instituciones internacionales como la *International Test Comisión, American Psychological Association, National Council on Measurement in Education o American Educational Research Association*.

En los últimos años son muchos los trabajos que se han dedicado a diseñar y mejorar procedimientos de detección del funcionamiento diferencial del ítem (Berk, 1982; Camilli y Shepard, 1994; Holland y Thayer, 1988; Millsap y Everson, 1993; Potenza y Dorans, 1995; Sheally y Stout, 1993). Sin embargo esta abundancia instrumental no ha ido pareja de trabajos cuyo objetivo haya sido la búsqueda de las causas de este error sistemático. Aún así, contamos con investigaciones relevantes sobre las causas del FDI en la adaptación de tests (Allalouf, Hambleton y Sireci, 1999; Ercikan, 2002; Elosua y López-Jaúregui, en prensa; Gierls y Khaliq, 2001; Hulin, 1987; Hulin y Mayer, 1986; van der Vijver y Tanzer, 1997). Estos autores hipotetizan varias fuentes de funcionamiento diferencial que podrían resumirse en los siguientes apartados: relevancia cultural, diferencias curriculares, diferencias gramaticales sean morfosintácticas o semánticas, entre idiomas o especificidades culturales que no se tienen en cuenta durante el proceso de adaptación y que causan una alteración en la

dificultad o poder discriminativo del ítem. La relevancia de cada una de estas potenciales fuentes de error dependerá de la diferencia entre los idiomas/culturas entre los que se lleva a cabo la adaptación. Así en entornos lingüísticos bilingües, similares a los que disponemos en nuestras comunidades autónomas, la adaptación de tests y cuestionarios, tendría que focalizar su atención en aspectos como las diferencias gramaticales o la idoneidad, adecuación y familiaridad del léxico utilizado.

En este marco de estudio del funcionamiento diferencial del ítem que integra tanto la detección como la comprensión, el objetivo de este trabajo es el análisis de los ítems liberados de la prueba de comprensión lectora que forma parte de la evaluación internacional de la OCDE PISA2000. El motivo de analizar solamente los ítems liberados es que queremos estudiar cuales son las razones que hayan podido generar FDI y para ello es imprescindible llevar a cabo un minucioso análisis de contenido.

Las muestras comparadas son la muestra original, Reino Unido, y la muestra española. La elección del Reino Unido queda justificada por ser uno de los países de referencia en la construcción de la prueba.

2. MÉTODO

Prueba

La prueba de comprensión lectora del programa PISA 2000 esta compuesta por 141 ítems divididos en 9 cuadernillos diferentes. El diseño de creación del programa de evaluación exige que todos los cuadernillos tengan un mínimo de ítems en común con el objetivo de poder equiparar las pruebas. Cada estudiantes participante en el proyecto responde a un sólo cuadernillo.

En este trabajo se han analizado los ítems liberados pertenecientes a los cuadernillos 8 y 9. Su rango de puntuaciones es 37 (0-37). El hecho de que los ítems liberados de ambos cuadernillos sean los mismos 26 ítems, hace posible utilizar un diseño de validación cruzada que permita legitimar las conclusiones del trabajo. Es decir, se llevan a cabo análisis independientes en cada uno de los cuadernillos con el fin de controlar las posibles falsas detecciones.

La estructura del test de comprensión lectora en PISA2000 es la siguiente: Se presenta a cada alumno un texto del que derivan una serie de preguntas a las que el alumno tiene que responder. Algunas de las preguntas son de opción múltiple, otras exigen una respuesta corta y en otras el alumno tiene que desarrollar su respuesta. Estas últimas se califican en una escala de respuesta graduada 0-2.

Los 26 ítems analizados están agrupados en 6 bloques de ítems correspondientes a 6 textos. De los 26 ítems a analizar 21 son de respuesta dicotómica y 5 son de respuesta graduada (0-1-2).

Participantes

La muestra de referencia esta formada por 2061 estudiantes pertenecientes al Reino Unido (1039 chicas, 1022 chicos) y la muestra focal es decir, la muestra española la forman 1350 estudiantes (676 chicas, 674 chicos). Todos los estudiantes tienen 15 años. La distribución de los alumnos por cuadernillos y países se recoge en la tabla 1.

Tabla 1. Estadísticos descriptivos por muestra y cuadernillo.

	Cuadernillo	N	Media Aritmética	Desviación típica	% varianza Primer Componente	α
Reino Unido	8	1032	20,82	8,88	29,8	0,913
	9	1029	21,45	8,88	29,4	0,899
España	8	675	17,51	7,71	22,5	0,861
	9	675	19,61	7,55	23,6	0,871

Análisis del funcionamiento diferencial del ítem

Se utilizan dos métodos inferenciales y un método descriptivo para la detección del funcionamiento diferencial del ítem. Los primeros ofrecen una prueba de significación sobre la existencia de funcionamiento diferencial del ítem, y el tercero proporciona un índice descriptivo del sentido e intensidad del FDI. Los índices inferenciales utilizados son el estadístico Mantel-Haenszel y la regresión logística que se adaptan para su aplicación a ítems politómicos. El índice descriptivo utilizado es la estandarización, que para el caso politómico se normaliza con el fin de obtener un indicador independiente del rango de puntuación del ítem.

Diferencia entre Medias Estandarizadas. (SMD) (Zwick y Thayer, 1996) Este índice es una extensión de la formulación de Dorans y Holland (1993) que proponen como indicador de FDI la diferencia entre las medias de los grupos de referencia y focal. El nuevo estadístico calcula la diferencia entre la media obtenida en el grupo focal y la media del grupo de referencia "estandarizada" como si la distribución del grupo de referencia fuera la misma que la del grupo focal. Un valor negativo en este índice indicaría que el ítem "favorece" al grupo de referencia.

Dado que el valor de este índice dependerá en el caso de ítems politómicos de la escala de respuesta, los autores proponen dividirlo por la desviación estándar del grupo focal y referencia combinados para obtener un índice que pueda ser interpretado independientemente de la escala de respuesta y que puede interpretarse como una medida del tamaño del efecto (SMD/S_i).

Mantel-Haenszel (1959) dicotómico. Es un procedimiento no-paramétrico para la evaluación de tablas de contingencia adaptado por Holland y Thayer (1988). Evalúa la igualdad entre las proporciones de respuestas correctas e incorrectas (odds-ratio) entre dos grupos (referencia y focal) a lo largo de los niveles en que se ha dividido la variable condicionante (puntuación total) por medio de un test Chi-cuadrado con 1 grado de libertad. La evaluación del tamaño del FDI se lleva a cabo con una transformación a la escala delta de los odds-ratio. Este indicador, delta mantel, indica la diferencia entre las dificultades entre el grupo de referencia y focal. Valores positivos indican que el ítem es diferencialmente mas fácil para el grupo focal. El *Educational Testing Service* (ETS) utiliza estos indicadores (significación del test de razón de proporciones y valor del estadístico delta) para catalogar la importancia del FDI. Un ítem presenta FDI severo cuando la prueba estadística es significativa y el valor absoluto del indicador Delta-Mantel es superior o igual a 1,5. La cantidad de FDI es moderada cuando la prueba estadística es significativa y el valor absoluto de Delta-Mantel está entre 1 y 1,5.

Mantel-Haenszel politómico. En el caso de ítems politómicos el procedimiento se adapta para el estudio de la distribución de las respuestas en cada una de las categorías de respuesta. (Spray y Miller, 1994; Tian, 1999; Zwick, Donogue y Grima, 1993; Elosua y López-Jaúregui, en prensa). La presencia del FDI se evalúa a través de un estadístico Chi-cuadrado con *m*-1 grados de libertad, siendo *m* es el número de categorías de respuesta. El tamaño del efecto puede analizarse a través de la diferencia entre medias estandarizadas (SMD; Dorans y Kulick, 1986) dividido por la desviación estándar de la combinación de los grupos de referencia y focal. Siguiendo el criterio utilizado por la ETS un ítem presente FDI moderado cuando además de la significación del estadístico utilizado, el tamaño del efecto es mayor o igual que 0,17 y menor o igual que 0,25. Por su parte el ítem presentara FDI severo si el tamaño del efecto es mayor que 0,25.

Regresión Logística. Este método modela la probabilidad de respuesta a un ítem en función de la puntuación empírica obtenida en el test, de la pertenencia a un grupo y de

la interacción entre ambos factores (Swaminathan y Rogers, 1990). Evalúa la presencia de FDI a través del estudio de la mejora en el ajuste que produce la incorporación sucesiva de los parámetros mencionados al modelo de regresión logística (Puntuación Total, Puntuación total+ Grupo, Puntuación total+Grupo+Puntuación total´Grupo). Para evaluar el FDI se comparan las razones de verosimilitud de los modelos anidados (total, total+grupo, total+grupo+interacción). El modelo base se construye únicamente respecto al parámetro de la variable que indica el nivel de habilidad (Total; Modelo 1). La existencia de FDI uniforme se concluiría cuando la diferencia entre el modelo base y el modelo que incluye el parámetro de pertenencia al grupo (Total+Grupo; Modelo 2) es significativa. El FDI no uniforme compara este segundo modelo con el modelo que incluye el término de interacción (Total+Grupo+Interacción; Modelo 3). Este método además de un test de significación incluye una medida del efecto del FDI basada en la diferencias en las R^2 entre dos modelos (Gierl y McEwen, 1998; Thomas y Zumbo, 1996). Esta medida, R^2, representa la proporción de variación de las respuestas al ítem explicada por el modelo de regresión. Un ítem presenta FDI moderado cuando el estadístico Chi-cuadrado es significativo y cuando el incremento en R^2 se sitúa entre los valores 0,035 y 0,070. Un ítem presenta un FDI notable cuando además de la significación del Chi-cuadrado, la diferencia entre dos R^2 es superior o igual a 0,070. Para las situaciones de respuestas politómicas el modelo se extiende dando lugar a tres variaciones básicas que dependen de la definición de los logit: el modelo acumulativo, el modelo continuo y el modelo de categorías adyacentes (Agresti 1984, 1990) siendo de todos ellos el más utilizado el modelo acumulativo. En este modelo se compara la probabilidad de que la respuesta al ítem (Y) sea menor o igual que la opción de respuesta *j* , con la probabilidad de que la respuesta (Y) sea mayor a la opción de respuesta *j*.

Los análisis se han llevado a cabo a través de un algoritmo implementado por la autora en S-Plus. La detección del FDI se ha efectuado en dos etapas. En la primera se ha detectado el FDI utilizando como variable condicionante todos los ítems; en una segunda fase se ha vuelto a estimar la variable condicionante (puntuación total) eliminando todos los ítems detectados en la etapa anterior (Holland y Thayer, 1988).

3. RESULTADOS

Resultados preliminares

Los estadísticos descriptivos para cada una de las muestras se presentan en la tabla 1. Comparaciones dentro de cada país: El test de Levene que evalúa la homogeneidad de las varianzas no ha sido significativo para ninguna de las dos comparaciones ($F_{LeveneSpain\ (1,1348)}$=1,43; p=0,23; $F_{LeveneReino\ Unido\ (1,2059)}$=0). Las diferencias de medias intrapaises y entre cuadernillos no son significativas en la muestra de referencia (t_{2059}=-1,61; p=0,108) y sí lo son en la muestra española (t_{1348}=-5,045; p<0,001). La media aritmética correspondiente al cuadernillo 9 (M.A.=19,61) es significativamente mayor que la media obtenida en el cuadernillo 8 (M.A.=17,51); el tamaño del efecto asociado a esta diferencia y estimado por eta cuadrado es 0,019.

Comparaciones dentro de cada cuadernillo. Las diferencias entre países para el mismo cuadernillo son significativas para los dos cuadernillos. Estas comparaciones se han llevado a cabo a través de la prueba t de Student bajo la condición de falta de homogeneidad de las varianzas ($F_{LeveneCuader.8\ (1,1705)}$=16,51; $F_{LeveneCuader.8\ (1,1702)}$=27,18). Tanto las diferencias relacionadas con el cuadernillo 8 (t_{1705}=-8,15; p<0,001) como las asociadas con el cuadernillo 9 (t_{1705}=-4,58; p<0,001) favorecen a la muestra de referencia. Las medidas del tamaño del efecto para cada una de las comparaciones estimadas con eta cuadrado son 0,035 y 0,011. En am-

bos casos las medias obtenidas por la muestra española son menores (ver tabla 1).

Unidimensionalidad y Consistencia Interna

Para cada uno de los cuadernillos analizados se ha extraído un componente principal en cada una de las cuatro muestras. La tabla 1 recoge el porcentaje de varianza asociado a cada uno de estos componentes. En la muestra de referencia, Reino Unido, el primer componente extraído explicó más del 29% de la varianza. En la muestra española este porcentaje es del 22,5% para el cuadernillo 8 y 23,6% para el cuadernillo 9.

El coeficiente alpha de Cronbach alcanza los valores de 0,899 y 0,913 en la muestra de referencia. Los valores obtenidos en la muestra española son 0,861 y 0,871 (tabla 1).

Funcionamiento Diferencial del Ítem

Los análisis se llevan a cabo de modo independiente en cada uno de los cuadernillos.

La tabla 2 muestra los resultados de los análisis realizados con lo ítems liberados pertenecientes al cuadernillo 8. Los resultados obtenidos con los ítems del cuadernillo 9 se muestran en la tabla 3. En ambas tablas puede leerse por columnas: el nombre del grupo de ítems analizado, el ítem analizado, ORD si el ítem es de respuesta graduada, y los valores obtenidos por cada uno de los estadísticos utilizados: Mantel-Haenszel, significación (p), Delta-Mantel, valor de la diferencia de medias estandarizada (SMD), diferencia de medias estandarizada normalizada (SMD/Si), los valores del ajuste de cada uno de los modelos de regresión logística utilizados a través del estadístico razón de verosimilitud (-2Log) y la diferencia entre las R^2 de Nagelkerke entre el modelo que incluye la puntuación total, la pertenencia la grupo y la interacción entre ambos (modelo 3) y el modelo base (modelo 1) que incluye en su especificación únicamente la puntuación total. En todos los casos se han marcado en negrilla los ítems con FDI.

Siguiendo los criterios expuestos para la clasificación de ítems con FDI para cada uno de los procedimientos utilizados podríamos resumir los resultados del siguiente modo:

Mantel-Haenszel: Este procedimiento detecta FDI en 9 ítems dicotómicos. De ellos 7 favorecen al grupo focal y 2 a la muestra inglesa o muestra de referencia. Del total de ítems con FDI 6 pueden ser calificados como severos. De los 5 ítems politómicos analizados 2 tienen un funcionamiento diferencial severo.

Regresión Logística: Este procedimiento detecta 4 ítems con funcionamiento diferencial moderado de los cuales 2 se refieren a ítems politómicos. Estos ítems han sido en su totalidad detectados por el procedimiento Mantel-Haenszel.

Los resultados del cuadernillo 9 pueden resumirse del siguiente modo.

Mantel-Haenszel. Este procedimiento detecta FDI en 11 ítems dicotómicos de los cuales 9 favorecen al grupo focal y 2 al grupo de referencia. Del total de ítems dicotómicos catalogados como FDI 6 son severos. De los 5 ítems ordinales analizados sólo uno presenta funcionamiento diferencial.

Regresión Logística: Este procedimiento arroja un total de 2 ítems con funcionamiento diferencial; uno de ellos es politómico y el otro es dicotómico. Estos dos ítems han sido también detectados por el procedimiento Mantel-Haenszel.

Tabla 2. Funcionamiento diferencial del ítem. Cuadernillo 8.

			Mantel	p	Delta-Mantel	SMD	SMD/Si	-2Log Modelo 1	-2Log Modelo 2	-2Log Modelo 3	Diferencia R^2 $Modo_3$-Mod_1
R040	*Q02*		3,47	0,06	-0,5	-0,05	-0,1	1800,9	1796,6	1782,8	0,012
	Q03A		**48,22**	**0**	**-2**	**-0,16**	**-0,32**	**1726,4**	**1669,1**	**1667,9**	**0,036**
	Q03B		**52,16**	**0**	**2,45**	**0,16**	**0,32**	**1381,3**	**1327,6**	**1327,6**	**0,035**
	Q04		1,41	0,23	0,36	0,02	0,05	1614,9	1613,4	1610,8	0,003
	Q06		0,29	0,58	-0,2	-0,01	-0,02	1915,4	1914,6	1914,6	0
R077	Q02		8,63	0	0,96	0,06	0,14	1372,2	1363,9	1363,1	0,006
	Q03	**ORD**	**42,7**	**0**		**-0,28**	**-0,32**	**2152,7**	**2105**	**2103,5**	**0,041**
	Q04		0,067	0,79	-0,1	-0,01	0,04	1922,8	1922,6	1922,6	0
	Q05	ORD	12,7	0		-0,15	-0,16	2727,2	2708,5	2705,2	0,019
	Q06		0,43	0,51	0,19	0,02	0,04	1876,5	1876,3	1875	0,001
R088	**Q01**		**38,42**	**0**	**1,82**	**0,14**	**0,302**	1697,4	1660	1660	0,024
	Q03	ORD	3,13	0,07		-0,05	-0,08	2592,1	2587,6	2584,3	0,005
	Q04	ORD	19,1	0,25		0,12	0,16	2649,2	2635,3	2635,3	0,01
	Q05		2,03	0,15	-0,5	-0,038	-0,084	1460,2	1457,6	1451,8	0,006
	Q07		0,07	0,79	0,09	0,001	0,002	1570,9	1570,9	1565,7	0,004
R110	**Q01**		**25,11**	**0**	**1,88**	**0,096**	**0,266**	1113,8	1086,6	1085,9	0,023
	Q04		0,003	0,95	-0	0,005	0,014	1103	1103	1095,2	0,007
	Q05		**12,79**	**0**	**1,19**	**0,085**	**0,215**	1264,4	1249,6	1242	0,018
	Q06		**27,5**	**0**	**1,81**	**0,106**	**0,27**	1353,2	1322,9	1318,2	0,028
R216	Q01		8,16	0	0,88	0,065	0,14	1533,8	1526,2	1526,2	0,006
	Q02		**16,17**	**0**	**1,36**	**0,102**	**0,2**	1317,2	1299,6	1298,4	0,014
	Q03		**19,75**	**0**	**-1,6**	**-0,108**	**-0,21**	1226,2	1205,5	1203,5	0,015
	Q04		**14,66**	**0**	**1,37**	**0,096**	**0,193**	1242,8	1227,5	1227,3	0,012
	Q06		1,04	0,31	0,31	0,019	0,039	1679,7	1679	1667,1	0,008
R236	Q01		0,07	0	-0,9	0	0	1704,2	1704	1692,1	0,008
	Q02	**ORD**	**11,8**	**0,02**		**0,17**	**0,2**	**1524,4**	**1505,5**	**1500,3**	**0,061**

Uno de los problemas del estudio del funcionamiento diferencial del ítem con datos empíricos es la imposibilidad de controlar las falsas detecciones o errores tipo I a los que ningún procedimiento de detección es ajeno (Elosua, López y Torres, 2000). Un modo habitual de "controlar" este efecto es la utilización conjunta de varios procedimientos de detección, que en este trabajo además reforzamos con un estudio de validación cruzada a través del análisis de los mismos ítems en dos muestras independientes. Utilizando una regla de decisión conservadora y fijando nuestra atención de modo independiente en los resultados de cada uno de los cuadernillos, podríamos concluir que 4 ítem presentan funcionamiento diferencial en el cuadernillo 8, y 2 ítems presentan funcionamiento diferencial en el cuadernillo 9. Si restringimos estos resultados con el estudio de validación cruzada, es decir, comparamos los resultados obtenidos en los dos cuadernillos y por los dos procedimientos inferenciales utilizados podríamos concluir con suma cautela que del total de 26 ítems analizados son 2 los que han sido detectados por todos los procedimientos utilizados y en los dos cuadernillos analizados. Esta cantidad supone el 7,6% de los 26 ítems analizados.

Estos ítems son el R236Q02 y R040Q03B. En ambos casos el ítem favorece al grupo focal o muestra española En la figura 1 se recogen las medias aritméticas de estos ítems estimadas en cada uno de los niveles en que se ha dividido la escala de habilidad, puntuación total (sólo se presentan los ítem de un cuadernillo por ser similares en ambos casos). El ítem R040Q03 padece de funcionamiento diferencial uniforme y el ítem R236Q02 presenta un funcionamiento diferencial no-uniforme a lo largo del continuo de habilidad. Las diferencias en el R^2 de Na-

gelkerke para este ítem entre el modelo que incorpora el termino de interacción (grupo, puntuación) y el modelo que incorpora al pertenencia al grupo ($R^2_{modelo3}$-$R^2_{modelo2}$) es 0,052 para el cuadernillo 8; mientras que la diferencia asociada al FDI uniforme ($R^2_{modelo2}$-$R^2_{modelo1}$) es 0,015. Curiosamente estos ítems han sido también detectados en un estudio que compara las muestras provenientes de Estados Unidos y de España (Elosua y Hambleton, en prensa).

Una vez detectados los ítems la siguiente fase a seguir sería la búsqueda de las causas que lo han originado. En esta etapa del trabajo han participado dos profesionales de la traducción /adaptación de textos entre el ingles y español, a los que se les ha pedido que analicen las posibles diferencias en el formato, estructura y contenido tanto de los ítems detectados como de los textos base de los que han surgido para buscar una explicación a estas diferencias.

Tabla 3. Funcionamiento diferencial del ítem. Cuadernillo 9

			Mantel	p	Delta-Mantel	SMD	SMD/Si	-2Log Modelo 1	-2Log Modelo 2	-2Log Modelo 3	Diferencia R^2 $Modo_3$-Mod_1
R040	*Q02*		5,96	0,01	-0,7	-0,06	-0,13	1748	1740	1735	0,009
	Q03A		**36,2**	**0,002**	**-1,65**	**-0,14**	**-0,29**	1890	1846	1845	0,029
	Q03B		**74,73**	**0**	**2,62**	**0,21**	**0,42**	**1679**	**1602**	**1602**	**0,052**
	Q04		**19,93**	**0**	**1,56**	**0,085**	**0,21**	1261	1243	1243	0,015
	Q06		2,11	0,146	0,39	0,038	0,08	1963	1962	1962	0
R077	**Q02**		**11,74**	**0,01**	**1,11**	0,068	0,16	1378	1369	1369	0,007
	Q03	ORD	13,13	0		-0,14	-0,17	2203	2203	2203	0,02
	Q04		3,01	0,08	-0,46	-0,04	-0,09	1954	1948	1945	0,006
	Q05	ORD	9,32	0		-0,13	-0,15	3001	2990	2989	0,007
	Q06		0,16	0,68	-0,12	-0,01	-0,02	1833	1832	1831	0
R088	**Q01**		**42,58**	**0**	**1,97**	0,147	0,33	1628	1586	1586	0,025
	Q03	ORD	0,169	0,68		-0,02	-0,02	2830	2828	2824	0,005
	Q04	ORD	7,63	0,01		0,087	0,13	2822	2818	2814	0
	Q05		0	-0,01	1	-0	-0,01	1410	1410	1409	0,003
	Q07		4,33	0,03	0,68	0,042	0,1	1435	1433	1428	0,007
R110	***Q01***		***14,55***	***0***	***1,42***	0,07	0,19	1107	1097	1096	0,02
	Q04		2,49	0,11	0,59	0,03	0,08	1096	1095	1091	0,006
	Q05		**13,91**	**0**	**1,32**	0,079	0,2	1140	1127	1125	0,007
	Q06		**18,24**	**0,003**	**1,43**	0,084	0,21	1347	1332	1331	0,008
R216	**Q01**		**30,76**	**0**	**1,96**	0,106	0,27	1271	1243	1243	0,03
	Q02		**22,36**	**0**	**1,51**	0,116	0,24	1470	1449	1449	0,003
	Q03		**11,95**	**0,001**	**-1**	*-0,08*	*-0,17*	1429	1416	1414	0,002
	Q04		0,47	0,49	0,22	0,016	0,03	1592	1591	1589	0,014
	Q06		2,87	0,09	0,54	0,036	0,08	1472	1470	1446	0,014
R236	Q01		1,32	0,25	-0,3	-0,03	-0,07	1600	1598	1597	0,003
	Q02	**ORD**	**16,08**	**0**		**0,18**	**0,21**	**1327**	**1310**	**1300**	**0,067**

El texto base del ítem 40 es un cuadro con información gráfica y datos estadísticos del que se derivan 5 preguntas. La pregunta que presenta funcionamiento diferencial es un ítem abierto en el que el alumno tiene que obtener información a partir del gráfico expuesto. La media del total de la población española que ha respondido a este ítem es 0,44 mientras que la media del resto de países participantes fue 0,36.

El texto base del ítem 236Q02 es el editorial de un periódico. Sobre él se han realizado dos preguntas abiertas de las cuales una presenta funcionamiento diferencial.

Ítem R040Q03B.

¿Por qué se ha elegido esta como fecha del comienzo del gráfico?

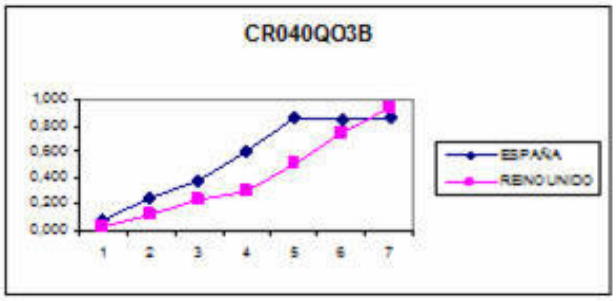
Why has the author chosen to start the graph at this point?

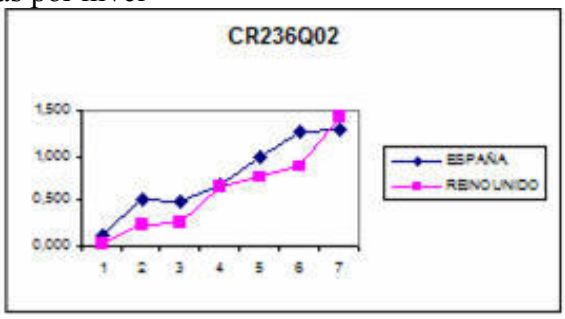
Item 236Q02

"List two examples from the editorial that illustrate how modern technology, such as that used for implanting frozen embryos, creates the need for new rules".

"Enumera dos ejemplos del editorial que justifiquen cómo la tecnología moderna, como la empleada para implantar embriones congelados, crea la necesidad de nuevas normas."

Figura 1. Medias aritméticas por nivel





El análisis llevado a cabo por los especialistas no pudo descifrar las causas de este desajuste. Respecto al ítem 40 los dos filólogos han acordado la falta de una equivalencia literal entre las dos versiones del ítem y tal vez la mayor cercanía de la versión española (curiosamente la versión adaptada) respecto al gráfico analizado, ya que en el mismo no hay ninguna referencia hacia el autor del mismo.

4. DISCUSION Y CONCLUSIONES

A pesar del esmerado y celoso desarrollo de los proyectos de evaluación internacional en todas sus fases (definición de objetivos, construcción de ítems, selección de muestras y proceso de adaptación de pruebas a distintos idiomas) la existencia de ítems que presentan funcionamiento diferencial sigue siendo un estigma permanente. Aunque el control de los errores tipo I a los que no son inmunes ninguno de los procedimientos de detección del FDI a través de la utilización de más de un procedimiento de detección (en nuestro caso Mantel-Haenszel, regresión logística y diferencia entre medias estandarizadas) y de la utilización como se ha hecho en este trabajo de una muestra de validación cruzada han permitido comprimir el número de ítems que presentan funcionamiento diferencial éste no ha sido eliminado. La consecuencia de este hecho es a todas luces negativa porque entre otros efectos podría cuestionarse el uso de las puntuaciones derivadas de estos ítems.

Aunque se está trabajando en el terreno de la detección temprana de funcionamiento diferencial a través de la aplicación de modelos teóricos que den cuenta de las causas del FDI todavía no se dispone de un regla de actuación generalizada que pueda prevenirlo (Elosua y Hambleton, en prensa). Para dar validez a un protocolo de actuación de este tipo habría de estudiarse de modo sistemático las causas que puedan producir un funcionamiento diferencial del ítem entre distintas lenguas/culturas para llegar a definir un marco ordenado que pueda explicar, predecir y eliminar los ítems con FDI. Aunque la elimi-

nación total de este error pueda ser ilusoria, sería positivo llevar a cabo estudios como los realizados en este trabajo con el fin de definir y delimitar para cada uno de los idiomas aquellas características más susceptibles de ser causa de FDI. En este sentido hemos de reconocer que el conjunto de ítems analizados en este trabajo es escaso con relación a los objetivos propuestos; sin embargo la exigencia propia del trabajo de disponer del contenido de los ítems junto a la necesidad de una muestra de validación han imposibilitado el acceso a un banco de ítems mayor.

Los resultados mostrados en este estudio han analizado el funcionamiento diferencial entre dos idiomas, el inglés y el español, y dos naciones, Reino Unido y España, que antes de la aplicación de las pruebas que componen la evaluación PISA han sido sometidas a un estudio de equivalencia exhaustivo. Si tenemos en cuenta que en las últimas evaluaciones internacionales (2003-2006) cada una de las Comunidades Autónomas aporta muestras propias, y que además, realiza la evaluación en los idiomas oficiales de cada una de ellas, la necesidad de los estudios mostrados en este trabajo se incrementa. Baste mencionar como ejemplo que un estudio sobre el funcionamiento diferencial asociado con la adaptación de pruebas (español/vasco) de rendimiento reveló la presencia de un porcentaje de FDI superior al 50% (Elosua, López y Egaña, 2000).

En definitiva la utilización conjunta de métodos empíricos de detección de FDI con análisis sustantivos que analicen en profundidad la estructura y contenido de los ítems marcaría la manera de proceder en aras a una detección temprana de FDI. Esta reciprocidad entre ambos tipos de métodos exige un análisis sistemático y previo entre adaptaciones. Los resultados de ese examen pueden dar lugar a un protocolo a seguir en el proceso de adaptación de tests que permita ahorrar tiempo y dinero a la par que incrementar la validez de las pruebas de evaluación (Elosua y Hambleton, en prensa).

5. REFERENCIAS

Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: Wiley and Sons.

Agresti, A. (1990). *Categorical Data Analysis.* New York: Wiley and Sons.

Allauf, A., Hambleton, R.K., y Sireci, S.G. (1999). Identifying the Causes of DIF in Translated Verbal Items. *Journal of Educational Measurement,* 36(3), 185-198

Berk, R. A. (Ed.) (1982). *Handbook of methods for detecting item bias*. Baltimore, John Hopkins Universtity Press.

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

Dorans, N.J., y Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and Standarization. En P.W.Holland y H.Wainer (Eds.) *Differential Item Functioning* (pp. 35-66) Hillsdale, NJ: Erlbaum

Dorans, N. J., y Kulick, E. (1986). Demonstrating the utility of the standarization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of educational measurement* **23**(4), 355-368.

Elosua, P., y Hambleton, R.K. (en prensa). Improving the Methodology for Detecting Biased Test Items. *International Journal of Testing*

Elosua, P., y López, A. (1999). Funcionamiento diferencial de los ítems y sesgo en la adaptación de dos pruebas verbales. *Psicológica, 20*, 23-40.

Elosua, P., y López-Jauregui, A. (en prensa). Potential DIF sources in the adaptation of tests. *International Journal of Testing.*

Elosua, P., y López-Jauregui, A. (en prensa). Aplicación de cuatro procedimientos de detección del funcionamiento diferencial sobre ítems politómicos. *Psicothema*.

Elosua, P., López, A., y Egaña, J. (2000). Idioma de aplicación y rendimiento en una prueba de comprensión verbal. *Psicothema* **12**(2), 201-206.

Elosua, P., López, A., Egaña, J., Artamendi, J. A., y Yenes, F. (2000). Funcionamiento diferencial de los ítems en la aplicación de

pruebas psicológicas en entornos bilingües. *Revista de Metodología de las Ciencias del Comportamiento, 2*(1), 17-33.

Elosua, P., López, A., y Torres, E. (2000). Desarrollos didácticos y funcionamiento diferencial de los ítems. Problemas inherentes a toda investigación empírica sobre sesgo. *Psicothema,* 12(2), 198-202.

Ercikan, K. (2002). Disentangling Sources of Differential Item Functioning in Multilanguage Assessments. *International Journal of Testing* **2**(3-4), 199-215.

Gierl, M.J., y McEwen, N. (1995). Differential Item Functioning on the Alberta Educationn Social Stuies 30 Diploma Exams. *Paper presented at the annual meeting of the Canadian Society for Studies in Education, Ottawa, Ontario, Canada.*

Gierl, M. J., y Khaliq, S.N. (2001). Identifying Sources of Differential Item and Bundle Functioning on Translaten Achievement Tests: A Confirmatory Analysis. *Journal of Educational Measurement* **38**(2), 164-187.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*(3), 164-172.

Hambleton, R. K., Merenda, P.F., y Spielberger, C.D. (Eds.) (2005). *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. Mahwah, NJ, Lawrence Erlbaum Associates.

Hambleton, R.K., y Jones, R.W. (1994). Comparison of Empirical and Judgmental Procedures For Detecting Differential Item Functioning. *Educational Research Quarterly,* 18 (1), 21-37.

Hambleton, R. K., y Patsula, L. (1999). Increasing the validity of Adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied testing Technology, 1*(1).

Holland, P.W., y Thayer, D.T. (1988). Differential Item Performance and the Mantel-Haenszel procedure. En H. Wainer y H.J. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum

Hulin, C.L. (1987). A psychometric theory of evaluations of Item Scale Translations. *Journal of Cross-Cultural Psychology*, 18(2), 115-142.

Hulin, C.L., y Mayer, L. (1986). Psychometric equivalence of a translation of the job descriptive index into Hebrew. *Journal of Applied Psychology*, 71(1), 83-94.

INECSE (2005). *Programa PISA. Pruebas de Comprensión Lectora*. Madrid: INECSE

López, A., y Elosua, P. (2002). Análisis de contenido y funcionamiento diferencial del ítem en una prueba de aptitud numérica. *Revista de Psicología General y Aplicada* **55**(3), 349-362.

Mantel, N., y Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.

Millsap, R. E., y Everson, H.T. (1993). Methodology review: statistical approaches for assessing measurement bias. *Applied psychological measurement* **17**(4), 297-334.

Potenza, M. T., y Dorans, N.J. (1995). DIF assessment for polytomously scored items: a framework for classification and evaluation. *Applied psychological measurement* **19**(1), 23-37.

Rogers, H.J., y Swaminathan, H. (1993) A comparison of the logistic regression and Mantel-Hanszel procedures for detecting Differential Item Functioning. *Applied Psychological Measurement,* 17(2), 105-117.

Shealy, R., y Stout, W. (1993). A model-based standarization approach that separates true bias/dif from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika* **58**(2), 159-194.

Spray, J., y Miller, T. (1994). Identifying nonuniform DIF in polytomously scored test items (American College Testing Research Report Series 94-1). Iowa City, IA: American College Testing Program.

Swaminathan, H., y Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of educational measurement* **27**(4), 361-370.

Tian, F. (1999). *Detecting differential item functioning in polytomous items*. Unpub-

lished doctoral dissertation, Faculty of Education, University of Ottawa.

Thomas, D.R., y Zumbo, B.D. (1996). Variable importance in regression and related analysis. Paper presented at the Annual Meeting of the Psychometric Society, Banff, AB, Canada.

van de Vijver, F. J. R., y Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47*(4), 263-279.

Zwick, R., Donogue J. R., y Grima, K.L. (1993). Assessment of Differential Item Functioning for Performance Tasks. *Journal of Educational Measurement* **30**(3), 233-251.

Zwick, R., y D. T. Thayer (1996). Evaluating the magnitude of differential item functioning in polytomous items *Journal of educational and behavioral statistics* **21**(3), 187-201.

NOTAS

Trabajo financiado por el Ministerio de Educación (SEJ2005-01694) y la Universidad del País Vasco. /1/UPV 00109.231-H-15904/2004)

ABOUT THE AUTHORS / SOBRE LOS AUTORES

Paula Elosua (paula.elosua@ehu.es) es profesora titular de Metodología de las Ciencias del Comportamiento en la Facultad de Psicología de la UPV/EHU. Su campo de especialización es la investigación psicométrica, especialmente la relacionada con las siguientes áreas: TRI, Validez, Adaptación de tests, Funcionamiento Diferencial del ítem, Equiparación de puntuaciones, SEM.

ARTICLE RECORD / FICHA DEL ARTÍCULO

Reference / Referencia	Elosua, Paula (2006). Funcionamiento diferencial del item en la evaluación internacional PISA. Detección y comprensión. *RELIEVE*, *v. 12*, n. 2. http://www.uv.es/RELIEVE/v12n2/RELIEVEv12n2_4.htm . Consultado en (*poner fecha*).
Title / Título	Funcionamiento diferencial del item en la evaluación internacional PISA. Detección y comprensión. [*Differential Item Functioning in the PISA Project: Detection and Understanding*]
Authors / Autores	Paula Elosua
Review / Revista	Revista ELectrónica de Investigación y EValuación Educativa (RELIEVE), *v. 12*, n. 2
ISSN	1134-4032
Publication date / Fecha de publicación	2006 (**Reception Date**: 2006 June 8; **Approval Date**: 2006 November 7; **Publication Date**: 2006 November 7)
Abstract / Resumen	*This report analyses the differential item functioning (DIF) in the Programme for Indicators of Student Achievement PISA2000. The items studied are coming from the Reading Comprehension Test. We analyzed the released items from this year because we wanted to join the detection of DIF and its understanding. The reference group is the sample of United Kingdom and the focal group is the Spanish sample. The procedures of detection are Mantel-Haenszel, Logistic Regression and the standardized mean difference, and their extensions for polytomous items. Two items were flagged and the post-hoc analysis didn't explain the causes of DIF entirely.* Este trabajo analiza el funcionamiento diferencial del ítem (FDI) de la prueba de comprensión lectora de la evaluación PISA2000 entre la muestras del Reino Unido y España. Se estudian los ítems liberados con el fin de aunar las fases de detección del FDI con la comprensión de sus causas. En la fase de detección se comparan los resultados de los procedimientos Mantel-Haenszel, Regresión Logística y Medias Estandarizadas en sus versiones para ítems dicotómicos y politómicos. Los resultados muestran que dos ítems presentan funcionamiento diferencial aunque el estudio post-hoc llevado a cabo sobre su contenido no ha podido precisar sus causas.
Keywords Descriptores	*Differential Item Functioning, PISA, Mantel-Haenszel, Logistic Regression, Polytomous DIF, Test adaptation* Funcionamiento Diferencial del ítem, PISA, Mantel-Haenszel, Regresión Logística, FDI politómico, Adaptación de tests
Institution / Institución	Universidad del País Vasco (España)
Publication site / Dirección	http://www.uv.es/RELIEVE
Language / Idioma	Spanish (Title, abstract and keywords in English)

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa
(**RELIEVE**)

[ISSN: 1134-4032]