Psicothema 2006. Vol. 18, nº 2, pp. 319-325
www.psicothema.com
ISSN 0214 - 9915 CODEN PSOTEG

Análisis de la variable género en las escalas del EDTC mediante técnicas de funcionamiento diferencial de los ítems

Sergio Escorial y María J. Navas*
C.E.S. Cardenal Cisneros y * Universidad Nacional de Educación a Distancia

El objetivo es analizar las diferencias de género en las escalas de una prueba de reciente creación, la Escala de Dificultades de Temperamento de Cantoblanco (EDTC), que mide los siguientes rasgos: ausencia de miedo, búsqueda de sensaciones e impulsividad. Las diferencias de género serán examinadas utilizando la tecnología del funcionamiento diferencial de los ítems, para ver si las diferencias observadas entre hombres y mujeres obedecen a diferencias reales en su personalidad o, por el contrario, se deben a artefactos en los ítems de estas escalas. Los métodos utilizados son la estandarización, la prueba χ^2 de Lord, el modelo DFIT, el SIBTEST y la regresión logística. Los resultados obtenidos sugieren que, pese a existir algunos ítems con funcionamiento diferencial, las diferencias de género observadas parecen deberse realmente a una diferencia en el constructo de personalidad y no a problemas del instrumento utilizado en su medición.

Analysis of the gender variable in the EDTC using differential item functioning techniques. The aim of this work is to analyze the gender differences in the scales of a recently constructed test: the so-called EDTC. This test measures the following traits: sensation seeking, fearlessness, and impulsivity. Gender differences will be studied using Differential Item Functioning (DIF) techniques, in order to determine whether these differences are true differences in the assessed dimensions or if, on the contrary, they are the result of a mere artefact of the measuring instrument used. The methods used to study DIF are standardization, SIBTEST, logistic regression, Lord's χ^2 test, and indices based on the DFIT model. Despite the fact that some items with DIF exist, the gender differences observed seem to be the result of true differences in the measured personality constructs and they don't seem to be artificially produced by a bias in the test items.

Lykken (1995) propuso un modelo teórico para explicar el desarrollo de las personalidades antisociales. Este modelo postula que en el comportamiento antisocial influyen factores genéticos que constituyen un factor de vulnerabilidad a la delincuencia persistente. Que dicha vulnerabilidad se manifieste en una personalidad y conductas antisociales depende en buena medida de la capacidad del entorno para socializar al individuo. Según Lykken, la vulnerabilidad genética se relaciona estrechamente con tres rasgos: la impulsividad, la ausencia de miedo y la búsqueda de sensaciones. Si éstos se manifiestan en grado extremo, entonces el sujeto estaría predispuesto al comportamiento antisocial y podría ser insensible a un esfuerzo socializador normal; en este caso, el autor habla de psicopatía. Cuando el papel principal lo juega el entorno (por ejemplo, una práctica familiar negligente), Lykken habla de sociopatía.

El marco teórico para los rasgos de impulsividad y ausencia de miedo lo proporciona la teoría motivacional de Gray (1987): los niveles altos de impulsividad están mediados por la actividad del Sistema Activador de la Conducta (BAS) y los de la ausencia de miedo por el Sistema Inhibidor de la Conducta (BIS). Un niño que nazca con un BIS poco activo y un BAS normal será vulnerable a la psicopatía primaria. Si el BAS es muy activo y el BIS normal, será vulnerable a la psicopatía secundaria. Por su parte, Zuckerman (1979) define la búsqueda de sensaciones como la tendencia de una persona a participar en actividades físicas o socialmente arriesgadas por el mero placer de realizarlas. Este rasgo se ha asociado, entre otras cosas, a deportes de riesgo, profesiones peligrosas, consumo de drogas, conducta sexual de alto riesgo y preferencias estéticas excéntricas; también se ha vinculado con la delincuencia (Aluja, 1991). Para Lykken, niveles altos de búsqueda de sensaciones predisponen a conductas arriesgadas y antisociales. En conjunto, la expresión elevada de los tres rasgos predispone, que no determina, al comportamiento antisocial en los adultos.

Para contrastar empíricamente la propuesta de Lykken, Herrero, Ordóñez, Salas y Colom (2002) construyeron la Escala de Dificultades de Temperamento de Cantoblanco (EDTC), que permite evaluar los tres rasgos considerados por Lykken. En ese estudio evaluaron a dos muestras (una de adolescentes y otra de reclusos) con una batería de tests compuesta por el EDTC, el Cuestionario de Personalidad de Eysenck revisado (Eysenck y Eysenck, 1997), la escala de Búsqueda de Sensaciones de Zuckerman (Zuckerman, 1979; Pérez y Torrubia, 1986) y la Escala de Impulsividad de Barrat (1985). Los resultados proporcionan evidencia clara acerca de

Fecha recepción: 1-3-05 • Fecha aceptación: 14-10-05
Correspondencia: Sergio Escorial
C.E.S. Cardenal Cisneros
División de Psicología
28006 Madrid (Spain)
E-mail: sergio.escorial@uam.es

la validez convergente de las escalas del EDTC. Sin embargo, en este estudio no se analizan las diferencias de género en estos tres rasgos, que tradicionalmente apuntan hacia niveles más altos en todos ellos en los varones (Barratt, 1985; Colom y Jayme Zaro, 2004; Eysenck y Eysenck, 1997; Pérez y Torrubia, 1986; y Zuckerman, 1979).

Las diferencias de género observadas en los tres rasgos del modelo de Lykken apuntan hacia una mayor vulnerabilidad de los varones al comportamiento antisocial, lo que puede tener consecuencias sociales muy claras. Por ejemplo, es esperable que haya más delincuentes varones que mujeres, algo totalmente confirmado por las estadísticas oficiales de la Dirección General de Instituciones Penitenciarias, que indican que cerca del 90% de la población reclusa son hombres.

Así las cosas, resulta clave la cuestión de dilucidar si las diferencias observadas entre varones y mujeres en estos rasgos obedecen a diferencias reales en su personalidad, reflejan una realidad que habla de diversidad o, por el contrario, son artefactos producidos por los propios ítems del test.

La tecnología del Funcionamiento Diferencial (FD) de los ítems constituye una herramienta muy valiosa para examinar a fondo las diferencias en la actuación en un test o escala de sujetos pertenecientes a distintos grupos. El estudio del FD permite distinguir entre diferencias reales (impacto) y ficticias (FD) en la actuación de distintos grupos de sujetos.

Un ítem presenta impacto cuando la probabilidad de elegir una determinada alternativa de respuesta difiere de un grupo a otro. Un ítem presenta FD cuando, para sujetos con *idéntico* nivel en la característica medida con el test, la probabilidad de un sujeto de elegir una opción de respuesta depende del grupo al que éste pertenezca. Esto es, en un ítem con FD la probabilidad de un sujeto de elegir una determinada alternativa depende de su nivel en la característica evaluada *y* de su grupo de adscripción. Habitualmente, se denomina *grupo focal* al grupo de interés, al grupo minoritario o socialmente desfavorecido y *grupo de referencia* al grupo con el que se va a comparar el de interés, normalmente, el grupo mayoritario. Cuando la probabilidad de elegir una opción de respuesta, para sujetos con el mismo nivel estimado en el rasgo, es sistemáticamente mayor en un grupo que en otro a lo largo de todo el continuo del rasgo, se habla de FD uniforme; en caso contrario, de FD no uniforme.

El objetivo principal de esta investigación es analizar las diferencias de género en los rasgos del modelo de Lykken. Para ello, se procederá a evaluar el impacto y seguidamente el posible FD de los ítems de la prueba EDTC con distintas técnicas de detección. Lo que se persigue, básicamente, es encontrar evidencia empírica para determinar si las diferencias observadas en los rasgos son diferencias genuinas en el rasgo latente que subyace a la medida o bien están producidas artificialmente por un sesgo en los ítems que componen una determinada escala de la prueba.

Método

Participantes

La muestra incluye 783 participantes, seleccionados mediante un sistema de asignación de cuotas de género y edad, con un 55% de varones y un 45% de mujeres y un porcentaje similar de sujetos (en torno al 17%) en cada uno de los seis grupos de edad considerados (<20, 20-29, 30-39, 40-49, 50-59 y >60). La edad media es de 40 años, con una desviación típica de 18 y un rango entre los 14 y los 88 años. Las medias y desviaciones típicas de hombres y mujeres en la variable edad son similares en las tres escalas consideradas. Los datos fueron recogidos por 75 evaluadores que tenían instrucciones de administrar la prueba EDTC a un hombre y una mujer en cada grupo de edad. El trabajo de campo duró tres semanas y la tasa de no respuesta fue del 13%.

Medidas

Se administró el EDTC (Herrero y col., 2002), prueba compuesta por 50 ítems que describen dos situaciones distintas debiendo elegir el sujeto aquella en la que preferiría encontrarse. Esta prueba evalúa los 3 rasgos del modelo de Lykken: búsqueda de sensaciones (16 ítems), ausencia de miedo (18) e impulsividad (16). Los coeficientes de fiabilidad obtenidos con los datos del presente estudio fueron .82, .69 y .80, respectivamente, resultando muy similares en el grupo de hombres y mujeres.

Análisis

Para la detección del impacto se examinó la relación entre las variables género y respuesta a cada uno de los ítems de las tres escalas anteriores. Para las escalas globales, el impacto se evaluó realizando una prueba de significación de diferencia de medias en muestras independientes. Además, se calculó una medida del tamaño del efecto (*d*, la diferencia de medias estandarizada) a nivel de ítem y de escala.

Para la detección del FD se emplearon las siguientes técnicas: estandarización, SIBTEST, estadístico χ^2 de Lord, estadísticos basados en el modelo DFIT y en la regresión logística.

Dado que el estadístico χ^2 y los basados en el modelo DFIT operan desde el marco de la Teoría de Respuesta al Ítem (TRI), antes de calcular ninguno de estos estadísticos era preciso comprobar que los datos obtenidos mostraban un ajuste razonable a alguno de los modelos TRI. Para ello, se siguieron las directrices propuestas por Hambleton y Swaminathan (1985), examinando el cumplimiento de los supuestos de los modelos considerados y evaluando la bondad de ajuste y la invarianza de los parámetros. El modelo que mejor ajuste presentó a los datos fue el modelo logístico de dos parámetros, si bien fue preciso eliminar un ítem en la escala de búsqueda de sensaciones (el 16) y en la de impulsividad (el 33) y tres ítems en la de ausencia de miedo (6, 32 y 50).

Regresión logística

Swaminathan y Rogers (1990) fueron los primeros en utilizar la regresión logística en la detección del DIF. Se trata de determinar si en la función matemática necesaria para predecir las respuestas a un ítem basta con introducir el nivel de habilidad de los sujetos (modelo sin FD) o, por el contrario, se debe incluir un término referido al grupo de pertenencia del sujeto en cuestión (modelo de FD uniforme) o un término que recoja la interacción entre el grupo de pertenencia y la habilidad del sujeto (modelo de FD no uniforme):

$$P(u_i = 1) = e^z / (1 + e^z)$$
$$Z = \beta_0 + \beta_1\theta + \beta_2 G + \beta_3(\theta * G)$$

donde:

u_i es la respuesta dada por el sujeto al ítem *i*;

θ es el nivel del sujeto en la característica que mide el test y definida en este estudio como la puntuación total del sujeto en la escala en cuestión del test;

G es la variable grupo de pertenencia para la que se quiere estudiar el FD.

Si el coeficiente β_3 es significativo, entonces el ítem presenta FD no uniforme. Si éste no es significativo y β_2 sí lo es, entonces se trata de FD uniforme. Finalmente, si sólo β_1 es significativo, entonces el ítem no funciona de forma diferente en los distintos grupos.

Zumbo y Thomas (1997) proponen combinar la significación estadística con una medida del tamaño del efecto para concluir si un ítem presenta o no FD, utilizando una estrategia de comparación de modelos. La significación estadística se establece a partir de la diferencia en el valor del estadístico χ^2 de los modelos comparados. Para la significación práctica, se considera que existe algún efecto cuando el incremento en el coeficiente de determinación (en el presente estudio se ha utilizado el de Nagelkerke) en los modelos comparados es de al menos .035, de modo que valores entre .035 y .070 denotarían un FD moderado y por encima de .070 un FD severo (Jodoin y Gierl, 2001). Ésta es la estrategia adoptada en este estudio.

Estandarización

Este procedimiento fue propuesto por Dorans y Kulick (1986) y proporciona como índice numérico para cuantificar el FD uniforme la diferencia en proporciones estandarizadas y para cuantificar el no uniforme la raíz cuadrada de esas diferencias ponderadas y elevadas al cuadrado:

$$STD\ P-DIF = \sum_{j=1}^{J} \frac{W_j}{\sum_{j=1}^{J} W_j}\left(P_{F_j} - P_{R_j}\right)$$

$$RMWSD = \sqrt{\sum_{j=1}^{J} \frac{W_j}{\sum_{j=1}^{J} W_j}\left(P_{F_j} - P_{R_j}\right)^2}$$

donde:

W_j es el factor de ponderación en el nivel de puntuación *j*, normalmente el número de sujetos del grupo focal en ese nivel de puntuación;

P_{Fj} y P_{Rj} son las proporciones de sujetos que responden al ítem en la dirección del rasgo en el nivel de puntuación *j* para el grupo focal y de referencia, respectivamente.

Sólo se dispone de prueba de significación estadística para el primer indicador, por lo que sólo ha sido éste calculado e interpretado en el estudio. Se ha utilizado como software *Dimensionality-Based DIF/DBF Package*, desarrollado por el *William Stout Institute for Measurement* (1990).

Estadístico χ^2 de Lord (1980)

Este estadístico contrasta la hipótesis nula de que los parámetros que definen la curva característica del ítem son iguales en el grupo focal y de referencia o, lo que es lo mismo, que el ítem no presenta FD:

$$\chi^2 = v\ \Sigma^{-1}\ v' \equiv \chi^2_n$$

donde:

v es un vector fila con las diferencias entre grupos en los parámetros del ítem estimados por el modelo TRI en cuestión;

Σ^{-1} es la inversa de la matriz de varianzas-covarianzas de *v*;

n es el número de parámetros del modelo.

Su valor ha sido determinado con el programa LINKDIF (Waller, 1998), que, además, realiza directamente la equiparación previa de las estimaciones obtenidas para los parámetros de los ítems en los grupos focal y de referencia con el método de la curva característica del test. Estas estimaciones se han obtenido con el programa MULTILOG (Thissen, 1991).

Modelo DFIT

El modelo DFIT (Differential Functioning of Items and Tests) constituye la aproximación más reciente y novedosa de las utilizadas en este estudio. Fue propuesto por Raju, van der Linden y Fleer (1995) en el marco de la TRI y, como su propio nombre indica, permite estudiar no sólo el funcionamiento diferencial de los ítems, sino también de la prueba en su conjunto. En particular, contempla dos estadísticos de funcionamiento diferencial para el ítem (CDIF y NCDIF) y uno de funcionamiento diferencial para el test (DTF):

$$CDIF_i = Cov\left(d_i, D\right) + \mu_{d_i}\mu_D$$

$$NCDIF_i = \sigma^2_{d_i} + \mu^2_{d_i}$$

$$DTF = \sigma^2_D + \mu^2_D = \sum_{i=1}^{n} CDIF_i$$

donde:

d_i es la diferencia en la probabilidad de responder en la dirección del rasgo al ítem *i* en los grupos focal y de referencia;

D es la diferencia entre las puntuaciones verdaderas esperadas en la prueba para el grupo focal y de referencia.

Para los índices DTF y NCDIF existen pruebas de significación estadística, no así para el CDIF. En este caso, para interpretar los valores obtenidos se procede de la siguiente manera: si el índice global no es estadísticamente significativo, se asume que ningún valor de CDIF lo es, pero si el DTF es significativo, entonces se eliminan los ítems que presenten un mayor valor absoluto de CDIF, uno a uno, hasta que el DTF deje de ser significativo. Cuando esto suceda se etiqueta los ítems eliminados como afectados de FD.

Asimismo, se ha considerado fundamental introducir algún criterio de significación práctica en la interpretación de todos estos estadísticos, se disponga o no de prueba de significación estadística asociada. La razón por la cual se ha tomado esta decisión es porque, en el primer caso, se trata de pruebas muy dependientes del tamaño muestral disponible, que hace que valores prácticamente nulos del estadístico NCDIF resulten estadísticamente significativos. En el segundo caso, se ha observado también que valores de CDIF muy próximos a cero conducen igualmente a ítems etiquetados como con FD. La estrategia que aquí se propone es la siguiente: señalar como ítems con posibles problemas de FD aque-

llos que el procedimiento iterativo anterior señale (CDIF) o los que indique la prueba de significación estadística (NCDIF), siempre y cuando el valor obtenido para el estadístico (CDIF o NCDIF) supere un determinado valor considerado como crítico. Estos valores críticos se obtuvieron en un estudio previo de simulación realizado por Navas y Asun (pendiente de publicación), en el que se estimó la distribución muestral de ambos estadísticos con un total de 4.000 valores para cada uno. Los valores críticos corresponden al percentil 99.5 de dicha distribución.

Para obtener todos estos índices y realizar las correspondientes pruebas de significación hemos utilizado una versión modificada del programa LINKDIF.

Sibtest

Este procedimiento permite detectar funcionamiento diferencial de los ítems y del test. En particular, se comparan las medias de los grupos focal y de referencia en un subtest formado por ítems de los que se sospecha pueden exhibir FD (subtest sospechoso), igualando a los sujetos en base a la puntuación obtenida en otro subtest compuesto por los ítems de la prueba que se considera que no presentan FD (subtest válido) y se calcula el siguiente estadístico, para el que se dispone de su correspondiente prueba de significación estadística:

$$\beta = \sum_{j=0}^{m} W_j \left(\overline{Y}_{R_j} - \overline{Y}_{F_j} \right)$$

donde:

W_j es el factor de ponderación empleado en el nivel de habilidad *j*, habitualmente definido como la proporción de sujetos en el grupo focal con puntuación *j* en el subtest válido;

$\overline{Y}_{R_j}$ *y* $\overline{Y}_{F_j}$ son las medias en el subtest sospechoso de los sujetos con puntuación *j* en el subtest válido, para el grupo de referencia y focal, respectivamente.

En la presente investigación se utilizó la información obtenida con los procedimientos anteriormente descritos para decidir qué ítems formaban el subtest válido y cuáles el subtest sospechoso. En particular, este último incluía los ítems identificados con FD al menos por dos procedimientos de detección. Los análisis fueron realizados con el programa *Dimensionality-Based DIF/DBF Package*.

Resultados

Impacto

La tabla 1 resume los resultados obtenidos al evaluar el impacto a nivel de ítem, indicando los que presentaban una relación estadísticamente significativa (p<.01) con la variable género. Asimismo, recoge qué grupo era el que puntuaba más alto, así como el rango del tamaño del efecto (d) de aquellos ítems detectados con un impacto significativo.

Estos resultados muestran que las escalas de búsqueda de sensaciones y ausencia de miedo presentan una proporción muy alta de ítems con impacto favorable a los hombres. Por su parte, en la escala de impulsividad se detectan dos ítems con impacto, uno a favor de las mujeres y otro a favor de los hombres. En consonancia con lo anterior, las escalas en las que aparecen diferencias significativas en las puntuaciones totales son la de búsqueda de sensaciones y ausencia de miedo, en ambos casos a favor de los varones (véase tabla 2).

Funcionamiento diferencial

Con el fin de ver si estas diferencias se mantienen o desaparecen cuando se compara a hombres y mujeres igualados previamente en su nivel de búsqueda de sensaciones (ausencia de miedo o impulsividad), se procedió a realizar los correspondientes análisis para detectar el posible FD. La tabla 3 recoge los resultados obtenidos con los métodos de detección anteriormente descritos, excepto el SIBTEST. Los ítems incluidos son aquellos que los análisis han revelado como estadísticamente significativos (p<.01), indicándose cuándo procede si el FD favorece al grupo de hombres (H) o mujeres (M).

Con la regresión logística se detecta FD en cinco ítems de la escala de búsqueda de sensaciones, pero solamente el 13 y el 25 se aproximan al nivel fijado para el FD moderado; el único ítem con

Tabla 1
Ítems con impacto significativo

Escala	Favorables a los hombres	Favorables a las mujeres	Rango de *d*
Búsqueda de sensaciones	2, 8, 13, 25, 28, 35, 40, 43, 46	–	.206 - .574
Ausencia de miedo	4, 5, 12, 15, 22, 24, 30, 36, 44	–	.188 - .565
Impulsividad	31	20	.225 - .249

Tabla 2
Estadísticos descriptivos en cada escala para hombres y mujeres, comparación de medias y tamaño del efecto

	Hombres			Mujeres				
Escala	N	Media	DT	N	Media	DT	Estadístico de contraste (p)	*d*
Búsqueda de sensaciones	434	5.48	3.71	349	3.81	3.32	t= 6.62 (p= .000)	.424
Ausencia de miedo	434	5.31	3.15	349	3.83	2.60	t= 7.05 (p= .000)	.495
Impulsividad	434	6.15	3.86	349	6.21	3.69	t= -.216 (p= .83)	–

FD no uniforme es el 25. En la escala de ausencia de miedo dos ítems presentan FD uniforme, siendo sólo en uno de ellos el tamaño del efecto moderado (el 24). Por último, la escala de impulsividad tiene tres ítems con FD uniforme, pero el tamaño del efecto es reducido en todos ellos.

Los resultados obtenidos con el método de la estandarización son totalmente coincidentes con los de la regresión logística en las escalas de búsqueda de sensaciones y ausencia de miedo. En el caso de la impulsividad, se detectan ahora cuatro ítems más con FD, dos a favor de los varones y otros dos a favor de las mujeres. En todos los casos se trata de FD uniforme, ya que sólo se calculó el estadístico correspondiente a este tipo de FD, por ser el único para el que se dispone de prueba de significación estadística.

Para determinar qué tipo de FD presentan los ítems cuando se trabaja con métodos basados en la TRI (prueba χ^2 de Lord e índices basados en el modelo DFIT) es preciso examinar las medidas exactas del área de Raju —calculadas también por el programa LINKDIF— y obtener las correspondientes representaciones gráficas de las funciones características de estos ítems en el grupo focal y de referencia. Como se puede observar en la tabla 3, de los cuatro ítems detectados en la escala de búsqueda de sensaciones tres fueron también identificados por la regresión logística y la estandarización (el 9, 13 y 41). Algo parecido sucede con la escala de ausencia de miedo: se detectan con problemas los dos mismos ítems que identificaron los procedimientos anteriores (el 11 y 24) y uno más (el 15), que muestra FD no uniforme. Respecto a la escala de impulsividad, la consistencia con los resultados de los dos métodos anteriores es elevada, detectándose de nuevo ahora un ítem más con FD no uniforme (el 14). Al evaluar el funcionamiento diferencial del test en cada una de las escalas se observó que la única con un funcionamiento diferencial significativo era la de ausencia de miedo.

La tabla 4 presenta la información más relevante de los análisis con el procedimiento SIBTEST. Los resultados obtenidos confirman que la inmensa mayoría de los ítems del subtest sospechoso presentan funcionamiento diferencial (las únicas excepciones son los ítems 29 y 45 de la escala de impulsividad) mayoritariamente uniforme (sólo el ítem 25 de la escala de búsqueda de sensaciones muestra FD no uniforme). Pese a ello, no se aprecia funcionamiento diferencial en el subtest sospechoso en ninguna de las tres escalas (véase la última columna de la tabla). La razón es que se produce el fenómeno de cancelación, al haber ítems en todas las escalas que favorecen en unos casos al grupo de mujeres y en otros al de hombres.

Si se observan conjuntamente las tablas 3 y 4 se pone de manifiesto una notable consistencia en los resultados obtenidos al utilizar distintos procedimientos. Así, de los 14 ítems que forman los subtests sospechosos de las tres escalas, 8 fueron detectados con FD al menos por cuatro de los cinco procedimientos y en 4 ítems se detectó la presencia de FD por medio de tres procedimientos. Sólo en 2 ítems el consenso se redujo a dos métodos de detección. En la tabla 5 se han incluido los enunciados de estos 14 ítems junto con los valores obtenidos en los estadísticos de detección de DIF utilizados, así como la probabilidad asociada a los mismos.

Discusión

El objetivo de este estudio era comprobar si las diferencias de género observadas en los rasgos que evalúa el EDTC son diferencias genuinas en las dimensiones medidas o están provocadas por un funcionamiento diferencial de los ítems que componen las diferentes escalas de la prueba.

Los resultados muestran que las escalas en las que se han encontrado diferencias significativas entre hombres y mujeres han sido la de búsqueda de sensaciones y ausencia de miedo, tanto en las puntuaciones globales como en un buen número de sus ítems. La dirección de las diferencias es consistente con la literatura: son los hombres los que obtienen puntuaciones más altas en búsqueda de sensaciones y ausencia de miedo.

Sin embargo, las diferencias habitualmente encontradas en impulsividad no aparecen en este estudio. Esta inconsistencia con la literatura previa podría tener que ver con la variable edad: las diferencias en impulsividad entre hombres y mujeres podrían no ser constantes a lo largo del ciclo vital. Así, mientras en la adolescen-

Tabla 3
Ítems con posibles problemas de funcionamiento diferencial detectados con distintos procedimientos

Escala	Regresión logística			Estandarización		χ^2			Modelo DFIT					
									CDIF			NCDIF		
	FD U		FD NU	FD U		FD U		FD NU	FD U		FD NU	FD U		FD NU
	H	M		H	M	H	M		H	M		H	M	
Búsqueda de sensaciones	8,13	9,41	25	8,13,25	9,41	13	9,41,7	–	–	–	–	–	–	–
Ausencia de miedo	24	11	–	24	11	24	11	–	24	–	15	24	–	–
Impulsividad	26,31	20	–	26,31,34,48	20,29,45	26,31,34,48	20,29,45	14	–	–	–	26	–	–

Tabla 4
Resultados obtenidos con el procedimiento SIBTEST

Escala	Nº de ítems en subtest válido	Ítems en el subtest sospechoso	FD uniforme a favor de:		FD no uniforme	FDT (p)
			H	M		
Búsqueda de sensaciones	10	8, 9, 13, 25, 41	8,13,	9,41	25	.527
Ausencia de miedo	13	11, 24	24	11	–	.529
Impulsividad	8	20, 26, 29, 31, 34, 45, 48	26,31,34,48	20	–	.271

cia y primera adultez son los hombres los que presentan puntuaciones más altas, en edades posteriores tales diferencias se podrían disipar. Colom y Jayme Zaro (2004) así lo constatan en una reciente revisión de trabajos. En la mayor parte de los estudios se trabaja con muestras con un rango de edad muy homogéneo, con sujetos próximos a la primera adultez, mientras que en la presente investigación la banda de edad considerada es muy amplia.

El estudio posterior del posible FD de los ítems de estas escalas ha puesto de manifiesto que las diferencias reveladas inicialmente por el análisis del impacto son, en líneas generales, diferencias que obedecen a la diversidad según género en la estructura de la personalidad y no tanto a problemas en los instrumentos utilizados para evaluarla.

En efecto, sólo han sido identificados consistentemente (con al menos cuatro de los cinco métodos de detección) 3 ítems de la escala de búsqueda de sensaciones (9, 13 y 41), 2 ítems de la escala de ausencia de miedo (11 y 24) y 3 ítems de la escala de impulsividad (20, 26 y 31). Además, al favorecer el FD en unas ocasiones a los hombres y en otras a las mujeres, se cancela su efecto al ser considerados conjuntamente en el subtest sospechoso, de forma que no existe funcionamiento diferencial del subtest sospechoso, pese a existir ítems con FD en todas las escalas.

En definitiva, en la prueba EDTC se detecta FD —predominantemente uniforme— en muy pocos ítems en los que, además, el tamaño del efecto es como mucho moderado, observándose asimismo un efecto de cancelación que se traduce en que la escala en su conjunto no funciona de forma diferente en el grupo de hombres y mujeres.

A la vista de los resultados obtenidos sería interesante responder a la pregunta de si estos ítems que funcionan diferencialmente son, al menos en parte, responsables de las diferencias encontradas entre hombres y mujeres al analizar las escalas globalmente. Para dar respuesta a esta cuestión se seleccionaron de cada escala los ítems del subtest válido empleado en el SIBTEST, esto es, los ítems de los que tenemos fundadas razones para suponer que no funcionan diferencialmente en función del género y que proporcionan, en principio, una medida válida de la característica evaluada. Seguidamente, se calculó en dicho subtest la puntuación media de hombres y mujeres para cada escala y se realizó la correspondiente comparación de medias.

Los resultados muestran que las tendencias no cambian cuando son eliminados de las escalas los ítems con FD, es decir, en aquellas escalas en las que existían diferencias significativas entre hombres y mujeres se siguen produciendo diferencias, mientras

Tabla 5
Ítems sospechosos de DIF según distintos estadísticos de detección[1]

Escala	Ítem	Regresión logística	Estandarización	χ^2	SIBTEST
Búsqueda de sensaciones	8.a. Ir en un coche a 200 Km/h	9.98	.092	–	.093
	.b. Ir al cine a ver una película	(.005)	(.003)		(.001)
	9.a. Ir a una exposición acerca de un tema que te interesa	22.69	-.189	23.42	-.171
	b. Acudir a una fiesta hasta altas horas de la noche donde no sabes qué te vas a encontrar	(.000)	(.000)	(.000)	(.000)
	13.a. Terminar un puzzle que tienes a medias	33.63	.172	10.77	.162
	b. Pasar una noche íntima con una persona desconocida	(.000)	(.000)	(.004)	(.000)
	25.a. Aprender a utilizar un arma de fuego	19.65	.107	–	.085
	b. Alquilar una película de vídeo que acaba de salir de estreno	(.000)	(.002)		(.010)
	41.a. Escuchar música en tu habitación	13.53	-.133	15.89	-.116
	b. Conocer gente en una discoteca	(.001)	(.000)	(.000)	(.002)
Ausencia de miedo	11.a. Perder la cartera con mucho dinero y la documentación	20.93	-.145	11.87	-.150
	b. Enzarzarte en una pelea con un desconocido	(.000)	(.000)	(.003)	(.000)
	24.a. Encontrarte en calle solitaria con un perro en actitud agresiva	29.03	.218	86.18	.187
	b. Estar con fiebre muy alta toda una semana	(.000)	(.000)	(.000)	(.000)
Impulsividad	20.a. Antes de hacer una cosa repasas mentalmente si tienes todo lo que necesitas	16.43	-.142	29.81	-.101
	b. Vas a comprar y a menudo compras más de lo que necesitas	(.000)	(.000)	(.000)	(.006)
	26.a. La gente piensa de ti que eres impredecible	10.89	.104	27.57	.106
	b. Tus hábitos de higiene y de alimentación son muy estrictos	(.003)	(.003)	(.000)	(.004)
	29.a. Estás hablando con alguien que no está de acuerdo contigo y le levantas la voz		-.106	13.26	
	b. En tu vida diaria te fijas unos horarios que sueles cumplir	–	(.002)	(.001)	–
	31.a. Conoces a alguien una noche y cuando te das cuenta le has contado cosas íntimas, de lo que luego te arrepientes	11.00	.090	10.07	.101
	b. Quedas en casa con unos amigos y pasas el día organizando los detalles	(.003)	(.001)	(.006)	(.001)
	34.a. Consigues algún dinero y te lo gastas enseguida	–	.092	9.46	.110
	b. No te gusta apostar en los juegos de azar		(.009)	(.008)	(.003)
	45.a. Vas a cortarte el pelo y siempre dejas que sea el peluquero quien elija tu peinado		-.108	16.68	
	b. Llevas una cuenta detallada de todos tus gastos	–	(.002)	(.000)	–
	48.a. Te ofreces rápidamente para hacer cualquier cosa aunque luego te arrepientas		.108	13.57	.095
	b. Tienes tu habitación ordenada porque te gusta	–	(.002)	(.001)	(.009)

[1] Se indica entre paréntesis la probabilidad asociada a cada estadístico. No se han incluido los resultados obtenidos con los índices basados en el modelo DFIT al no disponerse de pruebas de significación estadística para todos ellos

que en las que tales diferencias no existían éstas siguen sin aparecer. Los tamaños del efecto para las diferencias son ahora ligeramente más reducidos que en la escala original (.385 y .466 para las escalas de búsqueda de sensaciones y ausencia de miedo, respectivamente).

Estos resultados tendrían algunas implicaciones dentro del modelo teórico de Lykken. La primera sería que el grupo de hombres, en principio, podría ser más vulnerable al desarrollo de comportamientos antisociales debido a su mayor nivel en dos de los tres rasgos propuestos por Lykken (búsqueda de sensaciones y ausencia de miedo). Por consiguiente, el esfuerzo que se requiere para socializar al individuo sería, en líneas generales, mayor para los hombres que para las mujeres. Segundo, los hombres podrían ser más vulnerables que las mujeres a la psicopatía primaria, dado que éstos presentan puntuaciones más altas en ausencia de miedo y dado que los niveles elevados de ausencia de miedo están mediados por un sistema inhibidor de la conducta poco activo. Por otro lado, al no existir diferencias en impulsividad (mediada por la actividad del sistema activador de la conducta), hombres y mujeres no diferirían en su vulnerabilidad a la psicopatía secundaria. Estos resultados son congruentes con lo que se encuentra en la literatura especializada sobre la mayor prevalencia de varones en la psicopatía primaria y en el trastorno antisocial de la personalidad (Hare, 1993, 2001).

En suma, los resultados obtenidos en este estudio aportan evidencia favorable a la interpretación de las puntuaciones de la escala EDTC en términos de los constructos teóricos propuestos por el modelo de Lykken.

Referencias

Aluja, A. (1991). *Personalidad desinhibida, agresividad y conducta antisocial.* Barcelona: Promociones y Publicaciones Universitarias.

Barratt, E.S. (1985). Impulsiveness subtraits: arousal and information processing. En J.T. Spence y C.E. Itard (eds.): *Motivation, emotion and personality.* North Holland: Elsevier.

Colom, R. y Jayme Zaro, M.J. (2004). *La Psicología de las diferencias de sexo.* Madrid: Biblioteca Nueva.

Dorans, N.J. y Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the SAT. *Journal of Educational Measurement, 23,* 355-368.

Eysenck, H.J. y Eysenck, S.B.G. (1997). *Cuestionario revisado de personalidad de Eysenck (EPQ-R).* Madrid: TEA Ediciones.

Gray, J.A. (1987). *The psychology of fear and stress (2nd ed.)*. Cambridge, England: Cambridge University Press.

Hambleton, R.K. y Swaminathan, H. (1985). *Item response theory: principles and applications.* Boston: Kluwer-Nijhoff.

Hare, R.D. (1993). *Without conscience: the disturbing world of the psychopaths among us*. Nueva York: Simon & Schuster.

Hare, R.D. (2001). Psychopaths and their nature: some implications for understanding human predatory violence. En A. Raine y J. Sanmartin (eds.): *Violence and Psychopathy* (pp. 5-34). Dordrecht, The Netherlands: Kluwer Academic Publishing.

Herrero, O., Ordóñez, F., Salas, A. y Colom, R. (2002). Adolescencia y comportamiento antisocial. *Psicothema, 14* (2),340-343.

Jodoin, M.G. y Gierl, M.J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regresión procedure for DIF detection. *Applied Measurement in Education, 14,* 329-349.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Lykken, D.T. (1995). *The antisocial personalities.* Hillsdale, NJ: Earlbaum.

Pérez, J. y Torrubia, R. (1986). Fiabilidad y validez de la versión española de la escala de búsqueda de sensaciones (Forma V). *Revista Latinoamericana de Psicología, 18* (1), 7-22.

Raju, N.S., van der Linden, W.J. y Fleer, P.J. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19* (4), 353-368.

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

Thissen, D. (1991). *MULTILOG: multiple, categorial item analysis and tests scoring using item response theory.* Chicago: Scientific Software, Inc.

Waller, N. (1998). LINKDIF: an S-PLUS routine for linking item parameter and calculating IRT measures of differential functioning of items and tests. *Applied Psychological Measurement, 22* (4), 392.

William Stout Institute for Measurement (1990). Dimensionality-Based DIF/DBF Package: SIBTEST, POLYSIBTEST, CROSSING SIBTEST, DIFSIM, DIFCOMP. IRT-Based Educational and Psychological Measurement Software.

Zuckerman, M. (1979). *Sensation seeking: beyond the optimal level of arousal*. Hillsdale: Erlbaum.

Zumbo, B.D. y Thomas, D.R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia: Prince George, B. C.