anales de psicología
1999, vol . 15, nº 2, 331-343

Murcia (España). ISSN: 0212-9728

Estudio del funcionamiento diferencial de los ítems en una Escala de Habilidades Sociales para Adolescentes

Maria Dolores Hidalgo Montesinos*, Francisca Galindo Garre[1], Cándido José Inglés Saura[1], Guillermo Campoy Menéndez[2] y Beatriz Ortiz Soria

Universidad de Murcia

Resumen: Este trabajo tiene como objetivo estudiar el posible funcionamiento diferencial de los ítems (DIF) que componen la Escala de Habilidades Sociales para Adolescentes (EHSPA) como una parte del proceso de análisis de ítems. Este estudio implementa dos procedimientos de detección del DIF (estadístico de Mantel-Haenszel y Modelos Logit) para examinar si los ítems de la EHSPA funcionan de forma distinta en grupos igualados en función del género.
Palabras clave: Habilidades sociales, funcionamiento diferencial del ítem, sesgo del ítem, análisis de ítems.

Title: A study of differential item functioning in a social skills scale for Adolescents.
Abstract: The purpose of this study was to determine the differential item functioning (DIF) of the Interpersonal Difficulty Scale for Adolescents (EHSPA) as part of the item analysis process. This study used two different DIF detection procedures (Mantel-Haenszel statistic and Logit Models) to evaluate the extent to which items in ESHPA functioned differently for matched gender groups.
Key words: Social skills, differential item functioning, item bias, item analysis.

Introducción

Desde los comienzos de la Psicología como disciplina científica, la evaluación tanto de variables de personalidad como de actitudes psicológicas en adultos, adolescentes y niños ha sido uno de los temas que más interés ha generado. En este ámbito, los estudios referidos a comprobar las posibles diferencias individuales, en dichas variables psicológicas, en poblaciones segregadas en función de indicadores sociodemográficos (por ejemplo, género, raza y nivel socioeconómico) han centrado gran parte de la investigación, aunque suscitando cierta polémica. Hay que considerar que un instrumento de medición no debe estar afectado, en su función de medir, por las características del objeto de medida y, en el grado en que lo esté, la validez del instrumento estará seriamente dañada, siendo la objetividad de la medida un requisito imprescindible para cualquier test o escala psicológica. Este tema, como se ha señalado anteriormente, ha generado una extensa área de investigación, y el foco de dicha polémica se encuentra en la afirmación que los tests y cuestionarios psicológicos están sesgados, lo cual, equivale a decir que, en igualdad de condiciones, los miembros pertenecientes a un grupo (si consideramos la variable género, por ejemplo, mujeres) obtienen sistemáticamente puntuaciones menores (o mayores) en ese test que los miembros de otro grupo (varones). A pesar de que los trabajos encuadrados en este área son popularmente conocidos como estudios de sesgo, en la actualidad está cada vez más extendido el término de *funcionamiento diferencial del ítem* (DIF). Un ítem funcionará diferencialmente o presenta DIF cuando dos grupos comparables de sujetos, es decir, con un nivel idéntico respecto al atributo medido por el test, lo ejecuten de manera distinta. Normalmente, el grupo objeto de análisis se denomina grupo focal y el grupo que sirve como criterio de comparación se conoce como grupo de referencia. Los ítems de un test pueden presentar distintos tipos de DIF (Mellenbergh, 1982): Se denomina DIF *uniforme* o consistente cuando no existe interacción entre el nivel del atributo medido y la pertenencia a un de-

* **Dirección para correspondencia**: Mª Dolores Hidalgo Montesinos. Dpto Psicología Básica y Metodología. Universidad de Murcia. Campus de Espinardo (Edif. "Luis Vives"). Aptdo. correos 4021, 30080 Murcia (España). E-mail: mdhidalg@fcu.um.es

1 Becarios del Programa de Formación de Profesorado y Personal Investigador de la Universidad de Murcia.
2 Becario de la Fundación Séneca (Proyecto de Investigación PB/15/FS/97)

terminado grupo; en cambio, hablaremos de DIF *no uniforme* o inconsistente cuando se dé esta interacción, es decir, cuando la diferencia de las probabilidades de responder correctamente al ítem en los dos grupos no sea la misma a lo largo de todos los niveles del atributo.

La investigación en este campo intenta determinar si características tales como raza y género, que pueden ser ajenas a los sujetos en su funcionamiento cognitivo y psicológico, pueden tener un efecto sobre la medida de un rasgo psicológico. En instrumentos de medida que juegan cada vez más un papel prominente en tareas tales como el diagnóstico de la conducta disfuncional, la identificación de poblaciones de riesgo y la asignación a programas de tratamiento, es necesario explorar y comprender cómo las características demográficas de los sujetos, su bagaje cultural y social interactúa en el proceso de medida de variables de personalidad.

La evaluación del sesgo en los tests se ha centrado mayoritariamente en grupos étnicos, y en menor medida en grupos definidos por el género (Gómez e Hidalgo, 1997). En situaciones aplicadas de evaluación y diagnóstico psicológico resulta interesante conocer si los instrumentos de medida que estamos utilizando están sesgados en función de la variable género. En este trabajo, nos referiremos a un área como la de las habilidades sociales, fundamental en etapas de la vida como la adolescencia, en la cual el papel de los roles sexuales, aunque presumiblemente importante, no ha sido suficientemente aclarado. Así, a pesar de la existencia de toda una cultura popular que afirma que las mujeres se conducen de forma más habilidosa en situaciones sociales que los hombres, no hay datos consistentes en la literatura sobre el impacto de las diferencias de género en las situaciones sociales (Caballo, 1993).

El objetivo de este trabajo es evaluar la posible presencia de funcionamiento diferencial de los ítems, en la *Escala de Habilidades Sociales para Adolescentes* (Méndez, Martínez, Sánchez e Hidalgo, 1995), en función de la variable género y distinguirlo del posible impacto (diferencias en la puntuación en el test debidas a diferencias reales en habilidades sociales). Esta escala consta originalmente de 160 ítems que se refieren a las distintas áreas sociales en las que se desenvuelve habitualmente el adolescente (instituto, familia, amigos y calle) y a las distintas "personas-estímulo" con las que el sujeto se relaciona, teniendo en cuenta tanto sus características (sexo, edad, grado de conocimiento y nivel de autoridad) como su número (uno, varias o muchas). El sujeto debe responder siguiendo una escala tipo Likert de cinco puntos, desde cero ("Ninguna dificultad") hasta cuatro ("Máxima dificultad"). Las clases de respuestas que se evalúan son: dificultad o no para dar opiniones, para hacer cumplidos, para dar las gracias, para iniciar una conversación, para pedir información, para exigir derechos, dar quejas, pedir favores, para empatizar y para pedir disculpas. De los 160 ítems originales sólo se utilizan 88 que conforman una estructura factorial de cinco factores (Méndez, Inglés, Hidalgo y Martínez, 1998): Situaciones de calle (factor I), Relaciones familiares (factor II), Asertividad (factor III), Interacciones con iguales (factor IV) y Cortesía (Factor V). El factor "Situaciones de calle" incluye 24 ítems que reflejan aspectos relativos a las interacciones de los adolescentes con desconocidos y conocidos en la vía pública y a las relaciones en un ambiente social. El factor "Relaciones familiares" viene definido por 25 ítems que valoran la dificultad del adolescente en interacciones con familiares (padre, madre y hermanos). La subescala "Asertividad" incluye 14 ítems referidos a la dificultad del adolescente para manifestar sus derechos en diversos ámbitos sociales. El factor denominado "Interacciones con iguales" (13 ítems) se refiere a las relaciones del adolescente con amigos y compañeros del colegio de ambos sexos. Por último, el factor "Cortesía", incluye 12 ítems que evalúan las dificultades del adolescente para pedir perdón o aceptar disculpas, así como para dar las gracias o recibir agradecimiento. Aunque este instrumento se encuentra en fase de desarrollo se aplicó a una muestra amplia de adolescentes para analizar el posible funcionamiento diferencial de algunos de los ítems que lo componen. Considerando el análisis del DIF como una fase más del proceso de elaboración de cuestionarios.

Método

Sujetos

Tras realizar un muestreo aleatorio por conglomerados (representados por cada una de las zonas geográficas de la Región de Murcia), se seleccionó una muestra de 841 adolescentes, 417 varones y 424 mujeres, con edades comprendidas entre 16 y 18 años (media = 17,20; desviación típica = 0,58). Todos los sujetos eran alumnos del Curso de Orientación Universitaria (C.O.U.) o del 1º Curso del Segundo Grado de Formación Profesional.

Procedimiento

Aplicación de la prueba. El entrevistador llevó a cabo una entrevista con los jefes de los departamentos de orientación y/o con los jefes de estudios de los centros para explicar los objetivos de investigación, presentar el instrumento de evaluación que se iba a aplicar y solicitar su colaboración.

Una vez obtenido el correspondiente permiso, los sujetos completaron voluntariamente la escala de forma colectiva en el aula durante la hora asignada a la actividad de tutoría. El entrevistador procedió a la entrega de los ejemplares que incluían las instrucciones y los ítems de la prueba; a continuación, leyó en voz alta las instrucciones, aclarando cualquier duda que surgiera, pero procurando no influir en la respuesta de los sujetos y advirtiendo la importancia de no dejar ningún ítem en blanco.

Análisis del impacto

Para evaluar la presencia o ausencia de impacto en cada ítem del test se ha llevado a cabo un contraste de hipótesis acerca de la igualdad o desigualdad de las proporciones de éxito obtenidas en cada grupo (varones versus mujeres). La prueba estadística utilizada viene dada por la siguiente expresión (Ironson, 1982; Linn y Harnisch, 1981):

$$Z_i = \frac{p_{iR} - p_{iF}}{\sqrt{p_i(1-p_i)\left[1/N_R + 1/N_F\right]}}$$

que sigue una distribución normal y donde, p_{iR} es la proporción de éxito del grupo de referencia (R) en el ítem *i*, p_{iF} es la proporción de éxito obtenida en el grupo focal (F) para ese mismo ítem, N_R y N_F son respectivamente el número de sujetos en el grupo de referencia y en el grupo focal, y por último, p_i es la proporción de éxito obtenida en ese ítem para toda la muestra, que viene dada por:

$$p_i = \frac{N_R p_{iR} + N_F p_{iF}}{N_R + N_F}$$

Análisis del DIF

La evaluación del DIF se abordó utilizando dos procedimientos diferentes basados en las puntuaciones observadas en el test: el estadístico de Mantel-Haenszel y el análisis mediante el ajuste de modelos Logit.

Estadístico de Mantel-Haenszel (MH). Es uno de los más utilizados para detectar ítems con DIF, ya que presenta pocas dificultades tanto de cálculo como de interpretación, es intuitivamente más comprensible para profesionales con poco dominio de la estadística y no requiere de tamaños muestrales excesivamente grandes (Holland y Thayer, 1988; López Pina, Hidalgo y Sánchez-Meca, 1993; Mazor, Clauser y Hambleton, 1992).

Este procedimiento compara la ejecución en un ítem entre el grupo de referencia y el grupo focal a través de los distintos niveles de un determinado criterio de equiparación, normalmente la puntuación observada en el test. En el cálculo de este estadístico, el continuo de habilidad se divide en K intervalos de habilidad, y se construyen K tablas de contingencia 2 x 2 para cada ítem del test sobre el que vamos a evaluar el DIF. En cada una de estas tablas los sujetos son clasificados según el grupo de pertenencia (focal o referencia) y las posibles respuestas al ítem (si o no). La forma de esta tabla para un intervalo de habilidad dado aparece en la tabla 1.

Tabla 1: Tabla de contingencia bidimensional para el análisis del DIF.

	Respuesta al Item i		
Grupo	Acierto	Fallo	TOTAL
Referencia	A_k	B_k	n_{rk}
Focal	C_k	D_k	n_{rk}
TOTAL	m_{1k}	m_{0k}	T_k

Para evaluar el funcionamiento diferencial en un ítem de un test, se obtiene el valor de α (cociente de razones) para cada nivel de habilidad, es decir, para cada subtabla 2 x 2. Este índice expresa el cociente o razón entre la probabilidad de acertar el ítem en el grupo de focal versus la probabilidad de fallarlo frente a la probabilidad de acertar el ítem en el grupo de referencia versus fallarlo en dicho grupo. Mantel y Haenszel (1959) proponen como estimador de α la siguiente expresión:

$$\hat{\alpha} = \frac{A_k D_k / T_k}{C_k B_k / T_k}$$

que puede adoptar valores entre 0 y ∞. Cuando α es igual a 1, no hay diferencias entre los grupos sometidos a evaluación, por lo que el ítem no presenta DIF. Sin embargo, cuando $\alpha > 1$, nos encontramos ante un ítem que favorece al grupo de referencia sobre el grupo focal; y si es menor de 1, el ítem favorece al grupo focal sobre el de referencia, es decir, es más fácil para el grupo focal. La hipótesis nula de no DIF se somete a comprobación mediante el siguiente estadístico (Holland y Thayer, 1988):

$$MH = \frac{\left(\left|\sum_{k=1}^{K} A_k - \sum_{k=1}^{K} E(A_k)\right| - 0.5\right)^2}{\sum_{k=1}^{K} Var(A_k)}$$

$$Var(A_k) = \frac{N_{Rk} N_{Fk} m_{1k} m_{0k}}{T_k^2 (T_k - 1)}$$

donde si la hipótesis nula es cierta, este estadístico sigue una distribución χ^2 con un grado de libertad.

Como limitación importante del estadístico de *MH* se encuentra su escasa potencia estadística para detectar el DIF no uniforme (Mellenbergh, 1982; Rogers y Swaminathan, 1993; Swaminathan y Rogers, 1990; Uttaro y Millsap, 1994). El procedimiento *MH*, aplicado de forma estándar, no resulta apropiado para detectar DIF no uniforme, en este caso los modelos Logit (también los modelos Loglineales y los de Regresión Logística, entre otros) son una buena alternativa.

Modelos Logit. El segundo procedimiento, utilizado en este trabajo, para detectar la presencia de ítems con DIF, se basa en el análisis y ajuste de modelos Logit (Bishop, Fienberg y Holland, 1975; Agresti, 1990; Hidalgo, 1995; Gómez e Hidalgo, 1997). La aplicación de estos modelos se fundamenta en el estudio de la distribución de las frecuencias dentro de una tabla de contingencia con *m* dimensiones, siendo *m* el número de variables. La situación más habitual en el estudio del DIF es una tabla tridimensional en la que se representan, el grupo, la respuesta al ítem y la puntuación en el criterio de equiparación (normalmente, y al igual que en el estadístico *MH*, este criterio suele ser la puntuación observada en el test). Una vez construida la tabla de contingencia es posible formular distintas hipótesis acerca de cómo se distribuyen las frecuencias en cada una de las celdillas de la tabla, es decir, acerca de la presencia o ausencia de DIF, así como del tipo de DIF (uniforme o no uniforme). Esta última posibilidad, hace de los modelos Logit una técnica especialmente ventajosa en comparación al estadístico *MH*, ya que es posible detectar DIF no uniforme. Sin embargo, también hay que señalar que el ajuste de estos modelos es más costoso, su análisis e interpretación requiere de un mayor dominio estadístico y su comprensión es menos intuitiva para profesionales de la educación y la psicología.

En el presente estudio, las variables independientes consideradas en el modelo fueron la puntuación observada del sujeto en el test (H) y el género (G). La variable dependiente fue la respuesta al ítem con dos niveles, donde se trabaja con una transformación logit del tipo ln *[p/1-p]*, siendo *p* la proporción de éxito. Los modelos de interés en el estudio del DIF son:

$$\ln(p_{kl1}/p_{kl2})=\eta+\eta^{H}_{k} \quad M(2)$$
$$\ln(p_{kl1}/p_{kl2})=\eta+\eta^{H}_{k}+\eta^{G}_{l} \quad M(3)$$
$$\ln(p_{kl1}/p_{kl2})=\eta+\eta^{H}_{k}+\eta^{G}_{l}+\eta^{HG}_{kl} \quad M(4)$$

siendo η el efecto total de la dificultad del ítem, η^{H}_{k} el efecto principal de la variable habilidad de los sujetos (agrupada en intervalos), η^{G}_{l} el efecto principal de la variable grupo y η^{HG}_{kl} el efecto que expresa la interacción entre habilidad y pertenencia a grupo. En el análisis del DIF si el modelo M(3) se ajusta a los datos, el ítem funciona diferencialmente, pero debemos determinar el tipo de DIF. Si es necesario el parámetro de interacción η^{HG}_{kl} para explicar el comportamiento de los datos, el ítem presentará DIF no uniforme. Por el contrario, si el modelo M(3) se ajusta adecuadamente a los datos (la diferencia entre el modelo M(2) y el modelo M(3) es significativa), el ítem estará uniformemente sesgado; es decir, para todos los intervalos de habilidad evaluados, las diferencias logit entre los dos grupos son constantes. Por último, si el modelo M(2) se ajusta adecuadamente a los datos, diremos que el ítem no presenta DIF.

En el ajuste y estimación de los distintos modelos se ha utilizado el programa GLIM (Francis, Green y Payne, 1993); para probar el ajuste de cada uno de los modelos citados anteriormente se ha trabajado con el estadístico G^2 o razón de verosimilitud (Bishop, Fienberg y Holland, 1975) donde se compara las frecuencias esperadas, bajo el supuesto de que el modelo sea correcto, con las frecuencias observadas. Este estadístico sigue una distribución χ^2 con los grados de libertad asociados al modelo que estemos ajustando. A efectos de seleccionar el modelo más adecuado se ha empleado el estadístico G^2 condicional de comparación de modelos (De Maris, 1991, 1992), que resulta especialmente útil para evaluar si la inclusión de un término en un modelo es o no significativa. Este estadístico adopta la siguiente expresión:

$$G^2(MC/MS) = G^2(MC) - G^2(MS)$$

donde MC es el modelo que incluye el término a probar y MS es el mismo modelo pero excluyendo dicho término. Este estadístico sigue una distribución χ^2 con grados de libertad igual a los grados de libertad del modelo más complejo (MC) menos los grados de libertad del modelo más sencillo (MS).

Dado el elevado número de ítems del cuestionario de habilidades sociales (n = 88) y el número de categorías de respuesta (categorías = 5), el tamaño muestral seleccionado resultó insuficiente para poder obtener resultados fiables acerca del posible funcionamiento diferencial de los ítems a través de la variable género. Las frecuencias para muchas de las categorías de respuesta eran cercanas a cero, la ausencia de información en algunas de las categorías de respuesta hace inapropiado el uso de procedimientos de detección del DIF para ítems politómicos y requiere del uso de los procedimientos para ítems dicotómicos. Por este motivo, las respuestas a los ítems fueron dicotomizadas de tal modo que se asignó una puntuación de 0 (ninguna o casi ninguna dificultad) cuando los sujetos puntuaban en el ítem dos puntos o menos y una puntuación de 1 (dificultad) cuando los sujetos puntuaban tres o cuatro en el ítem. A partir de estos datos trasformados, se realizaron todos los análisis y estudios del DIF. Además, tanto el análisis del impacto como el de DIF fueron realizados, para cada uno de los ítems del test, utilizando como criterio de equiparación la puntuación observada en cada uno de los factores.

Resultados

Análisis del impacto

La tabla 2 recoge los valores obtenidos para el estadístico Z_i en cada uno de los factores de la escala de habilidades sociales. Este estadístico de contraste somete a prueba la hipótesis de igualdad de proporciones de éxito para varones y para mujeres. Un valor positivo de Z_i indica que la proporción de éxito para los varones ("máxima dificultad") en ese ítem es mayor que la de las mujeres. Por el contrario, un valor negativo indica que la proporción de éxito es mayor para las mujeres que para los varones. Tal y como podemos observar en la tabla 2, para el factor I los ítems 20 y 24 evidencian impacto a un nivel de significación del 5% y el ítem 5 a un nivel de probabilidad me-

nor del 0.001, siendo el grupo de mujeres el que presenta una "mayor dificultad" que el grupo de varones.

Para el factor II sólo presentó un resultado estadísticamente significativo el ítem 16. En el factor III mostraron diferencias estadísticamente significativas seis de los 14 ítems que lo componen, en concreto los ítems 3, 7, 8, 9, 11 y 13 (ver tabla 2 para nivel de significación). En los ítems 4, 6-8 y 10-13 del factor IV los hombres tuvieron "mayor dificultad" que las mujeres. Por último, en el factor V sólo se encontraron diferencias significativas entre varones y mujeres en el ítem 11.

Análisis del DIF

Estadístico de Mantel-Haenszel. En la aplicación del procedimiento de Mantel-Haenszel, se utilizó como variable de equiparación la puntuación total observada en cada factor. Esta puntuación fue dividida en dos intervalos de habilidad, de tal modo que cada intervalo contuviera un porcentaje similar de sujetos y no hubiera en la tabla de contingencia celdillas con frecuencias nulas (Donoghue y Allen, 1993), el criterio de división fue la mediana.

La tabla 3 presenta los valores de *MH* para cada uno de los cinco factores de la escala de habilidades sociales.

Tabla 2: Resultados obtenidos en el análisis del impacto, en cada uno de los cinco factores.

FACTOR I		FACTOR II		FACTOR III		FACTOR IV		FACTOR V	
Item	**Zp**	**Item**	**Zp**	**Item**	**Zp**	**Item**	**Zp**	**Item**	**Zp**
1	-1.0547	1	-1.5466	1	-1.0439	1	-0.9780	1	-0.0811
2	.0001	2	-0.2474	2	-1.001	2	-1.6092	2	1.2274
3	-.3578	3	0.5479	3	-2.2089*	3	-1.3131	3	0.5700
4	-.6912	4	-1.4717	4	-1.1112	4	2.0680*	4	0.5993
5	-8.3767***	5	-1.2232	5	-0.5239	5	1.2178	5	-1.5836
6	-0.2454	6	-0.7050	6	0.1925	6	2.0302*	6	0.5482
7	-1.5394	7	-0.9398	7	-2.7322**	7	2.9488**	7	1.5933
8	-0.7837	8	-0.0511	8	-1.8834*	8	1.8929*	8	0.2134
9	-0.4998	9	-0.7436	9	2.4701**	9	1.4358	9	1.2396
10	-1.2384	10	-0.6118	10	1.1867	10	3.4122***	10	-0.2744
11	-1.2235	11	1.4190	11	-1.8842*	11	2.0691*	11	1.7977*
12	-1.1799	12	0.8975	12	-1.4712	12	3.9595***	12	0.0576
13	-0.8550	13	-0.8967	13	-1.9284*	13	2.3350**		
14	-1.4874	14	-0.9166	14	-0.9340				
15	-0.6949	15	-0.5406						
16	0.3911	16	-1.8661*						
17	-1.4294	17	0.4637						
18	-0.9280	18	0.2127						
19	-0.7626	19	-1.0522						
20	-2.1551*	20	-0.3545						
21	-0.6958	21	-0.7514						
22	-0.4860	22	-0.2584						
23	-1.2058	23	-0.4006						
24	-2.0594*	24	-0.5893						
		25	0.8696						

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

En el factor I tres de los 24 ítems obtienen valores del estadístico *MH* que resultaron significativos al 5%. En concreto estos ítems son el ítem número 16, 20 y el 24. La posibilidad de ítems que funcionan diferencialmente también se apunta en cuatro ítems del factor II y en cinco del factor III (ver tabla 3). En el factor IV siete de los 13 ítems resultaron sospechosos de DIF ya que la respuesta a los mismos depende tanto del nivel de habilidades sociales como de la variable género. Por último, en el factor V resultaron significativos cinco de los 12 ítems.

Modelos Logit. En la aplicación de los modelos Logit, también se utilizó como criterio de equiparación la puntuación total en el test dividida en dos intervalos. De este modo, obtuvimos 88 tablas tridimensionales ($2 \times 2 \times 2$), las cuales fueron sometidas a análisis. Las Tablas 4 a la 8 presentan los resultados obtenidos del ajuste de modelos Logit para cada uno de los factores de la escala de habilidades sociales. Los resultados se presentan detallados para cada ítem y para cada modelo ajustado. Así, en la segunda columna de la tabla 4 se muestran los valores de razón de verosimilitud para el modelo de no DIF (G^2 (M2)), siendo los grados de libertad (gl) asociados a este modelo igual a 2, en la tercera columna los valores de probabilidad asociados a dicho modelo, en la cuarta columna el valor de razón de verosimilitud para el modelo de DIF (G^2 (M3), gl=1), en la siguiente columna su correspondiente probabilidad y en las dos últimas columnas el estadístico de razón de verosimilitud condicional de comparación entre los dos modelos (ΔG^2, gl=1) y la probabilidad asociada al mismo.

Tabla 3: Valores del estadístico *MH* para cada uno de los cinco factores.

FACTOR I		FACTOR II		FACTOR III		FACTOR IV		FACTOR V	
Item	**MH**	**Item**	**MH**	**Item**	**MH**	**Item**	**MH**	**Item**	**MH**
1	1.497	1	5.681*	1	0.297	1	7.303**	1	0.006
2	3.177	2	0.004	2	0.187	2	15.418***	2	6.463*
3	0.322	3	2.128	3	6.165*	3	9.428**	3	0.946
4	0.037	4	4.848*	4	0.552	4	3.419	4	0.998
5	0.020	5	3.201	5	0.274	5	0.153	5	9.839**
6	0.504	6	0.643	6	2.045	6	3.584	6	11.485***
7	3.465	7	1.463	7	11.564***	7	9.223**	7	5.768*
8	0.118	8	0.110	8	3.505	8	2.808	8	0.132
9	0.369	9	0.774	9	36.976***	9	0.053	9	3.794
10	0.992	10	0.448	10	10.626***	10	13.344***	10	0.313
11	0.458	11	7.443**	11	3.386	11	3.666	11	8.679**
12	0.358	12	3.406	12	1.328	12	20.333***	12	0.006
13	0.275	13	1.278	13	4.156*	13	4.409*		
14	1.906	14	1.350	14	0.002				
15	0.041	15	0.340						
16	9.328**	16	6.845**						
17	1.618	17	1.119						
18	0.011	18	0.424						
19	0.022	19	1.897						
20	7.632**	20	0.102						
21	0.021	21	0.836						
22	0.160	22	0.024						
23	0.668	23	0.091						
24	5.775*	24	0.455						
		25	2.291						

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

Tabla 4: Factor I: Resultado del análisis del DIF con Modelos Logit.

Item	G^2 (M1)	p	G^2(M2)	p	ΔG^2	p
1	1.9449	0.3782	0.4424	0.2297	1.502	0.2204
2	3.4345	0.1796	0.2491	0.6177	3.185	0.0743
3	0.3469	0.8407	0.0237	0.8778	0.323	0.5696
4	1.5820	0.4534	1.5452	0.2138	0.037	0.8478
5	1.1022	0.5763	1.0820	0.2983	0.020	0.8871
6	5.0298	0.0809	4.5250	0.0334	0.505	0.4774
7	3.665	0.1600	0.1959	0.6581	3.469	0.0625
8	25.554	0.0000	25.548	0.0000	0.006	0.9361
9	2.5948	0.2732	2.2248	0.1358	0.370	0.5430
10	1.6224	0.4443	0.6296	0.4275	0.993	0.3191
11	3.7816	0.1510	3.3231	0.0683	0.458	0.4983
12	0.7030	0.7036	0.3446	0.5572	0.358	0.5493
13	0.2829	0.8681	0.0076	0.9303	0.275	0.5998
14	2.4015	0.3010	0.4944	0.4820	1.907	0.1673
15	1.3092	0.5196	1.2686	0.2600	0.041	0.8402
16	10.151	0.0062	0.7565	0.3844	9.395	0.0022**
17	2.0389	0.3608	0.4205	0.5167	1.618	0.2034
18	0.9894	0.6098	0.9789	0.3225	0.010	0.9182
19	0.0571	0.9718	0.0353	0.8510	0.022	0.8825
20	7.6573	0.0217	0.0243	0.8760	7.633	0.0057**
21	0.5239	0.7695	0.5025	0.4784	0.021	0.8836
22	0.1765	0.9156	0.0157	0.9003	0.161	0.6884
23	1.2533	0.5344	0.5836	0.4449	0.670	0.4132
24	7.1459	0.0281	1.3539	0.2446	5.792	0.0161*

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

Tabla 5: Factor II: Resultado del análisis del DIF con Modelos Logit.

Item	G^2 (M1)	p	G^2(M2)	p	ΔG^2	p
1	5.7163	0.0574	0.0054	0.9415	5.711	0.0168*
2	0.0339	0.9832	0.0299	0.8628	0.004	0.9491
3	2.9019	0.2343	0.7655	0.3816	2.136	0.1439
4	4.9980	0.0822	0.1303	0.7181	4.868	0.0274*
5	4.2322	0.1205	1.0185	0.3129	3.214	0.0730
6	1.7851	0.4096	1.1397	0.2857	0.645	0.4218
7	1.5100	0.4700	0.0411	0.8393	1.469	0.2255
8	1.2611	0.5323	1.1511	0.2833	0.110	0.7401
9	0.8223	0.6629	0.0459	0.8303	0.776	0.3782
10	0.7645	0.6823	0.3145	0.5749	0.450	0.5023
11	10.057	0.0065	2.5651	0.1092	7.492	0.0062**
12	5.2346	0.0703	1.8146	0.1780	3.420	0.0644
13	1.2842	0.5262	0.0006	0.9805	1.284	0.2572
14	5.9534	0.0509	4.5977	0.0320	1.356	0.2442
15	0.3662	0.8327	0.0250	0.8744	0.341	0.5591
16	10.392	0.0055	3.4241	0.0643	6.968	0.0083**
17	1.1457	0.5639	0.0236	0.8778	1.122	0.2895
18	0.5032	0.7776	0.0784	0.7795	0.425	0.5146
19	4.8006	0.0907	2.8932	0.0889	1.907	0.1673
20	1.4053	0.4953	1.3028	0.2537	0.103	0.7488
21	0.8410	0.6567	0.0001	0.9924	0.841	0.3591
22	0.2776	0.8704	0.2536	0.6146	0.024	0.8769
23	1.2256	0.5418	1.1340	0.2869	0.092	0.7622
24	2.0491	0.3590	1.5918	0.2071	0.457	0.4989
25	5.3917	0.0675	3.0887	0.7880	2.303	0.1291

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

Tabla 6: Factor III: Resultado del análisis del DIF con Modelos Logit.

Item	G^2 (M1)	p	G^2(M2)	p	ΔG^2	p
1	2.9522	0.2285	2.6552	0.1032	0.297	0.5857
2	6.1396	0.0464	5.9523	0.0147	0.187	0.6652
3	7.0994	0.0287	0.9013	0.3424	6.198	0.0128*
4	2.5545	0.2788	1.9998	0.1573	0.555	0.4564
5	0.4786	0.7872	0.2038	0.6517	0.275	0.6001
6	6.1837	0.0454	4.1370	0.0420	2.047	0.1525
7	13.198	0.0014	1.5514	0.2129	11.65	0.0006***
8	4.6988	0.0954	1.1692	0.2796	3.530	0.0603
9	42.739	0.0000	4.8381	0.0278	37.90	0.0000***
10	13.744	0.0010	3.0767	0.0794	10.67	0.0011**
11	8.4469	0.0147	5.0519	0.0246	3.395	0.0654
12	5.8847	0.0527	4.5515	0.0329	1.333	0.2483
13	4.7714	0.0920	0.6151	0.4329	4.156	0.0415*
14	0.3038	0.8591	0.3015	0.5829	0.002	0.9617

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

Tabla 7: Factor IV: Resultado del análisis del DIF con Modelos Logit.

Item	G^2 (M1)	p	G^2(M2)	p	ΔG^2	p
1	7.2037	0.0273	0.0001	0.9920	7.204	0.0072**
2	15.271	0.0005	0.0010	0.9748	15.27	0.0001***
3	9.3223	0.0095	0.0001	0.9920	9.322	0.0023**
4	3.5511	0.1694	0.0000	0.9999	3.551	0.0595
5	0.1535	0.9261	0.0000	0.9999	0.153	0.6953
6	3.7527	0.1531	0.0000	0.9999	3.753	0.0527
7	9.7340	0.0077	0.9737	0.9999	9.734	0.0018**
8	2.9182	0.2324	0.0000	0.9999	2.918	0.0876
9	0.0533	0.9737	0.0000	0.9999	0.053	0.8176
10	14.058	0.0009	0.0000	0.9999	14.06	0.0002***
11	3.8297	0.1474	0.0000	0.9999	3.830	0.0503
12	21.740	0.0000	0.0000	0.9999	21.74	0.0000
13	4.5603	0.1023	0.0000	0.9999	4.560	0.0327*

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

Tabla 8: Factor V: Resultado del análisis del DIF con Modelos Logit.

Item	G^2 (M1)	p	G^2(M2)	p	ΔG^2	p
1	1.9445	0.3782	1.9390	0.1638	0.006	0.9403
2	7.2546	0.0266	0.7553	0.3848	6.499	0.0108*
3	2.8193	0.2442	1.8699	0.1715	0.949	0.3299
4	2.6270	0.2689	1.6257	0.2023	1.001	0.3171
5	9.9507	0.0070	0.0264	0.8710	9.924	0.0016**
6	15.794	0.0004	4.2100	0.0402	11.58	0.0007***
7	6.0448	0.0487	0.1978	0.6565	5.847	0.0156*
8	0.2176	0.8969	0.0856	0.7699	1.132	0.7164
9	4.0186	0.1341	0.2009	0.6540	3.818	0.0507
10	0.5349	0.7653	0.2210	0.6383	0.314	0.5752
11	8.8633	0.0119	0.1216	0.7273	8.742	0.0031**
12	0.0077	0.9961	0.0014	0.9702	0.006	0.9364

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

En el factor I el ítem 8 no se ajustó ni al modelo M(2) ni al modelo de DIF (M3), ya que en ambos modelos la probabilidad asociada fue menor de 0.05. Estos resultados parecen indicar que para dicho ítem la interacción HG resultaría significativa, y por lo tanto necesaria para explicar la distribución de los datos, indicando DIF no uniforme. Como podemos observar en la tabla 4, para los ítems 16, 20 y 24, el modelo de DIF se ajustó substancialmente mejor que el modelo de no DIF, y además, el componente de DIF fue estadísticamente significativo. Las respuestas a estos ítems, al menos en parte, son función del género.

En la tabla 5 observamos que para los ítems número 1, 4, 11 y 16 del factor II, el modelo que mejor se ajusta a los datos es el modelo de DIF, y en concreto, que éste se presenta de modo uniforme. En el resto de ítems de éste factor, el modelo de no DIF es el que mejor representa los datos.

En el factor III los ítems 2, 6 y 11 no se ajustaron ni al modelo de no DIF ni al modelo de DIF. Tal y como se muestra en la tabla 6 los valores de G^2 para ambos modelos fueron bajos y su probabilidad asociada menor de 0.05. Estos resultados, al igual que ocurría en el ítem 8 del factor I, parecen indicar que dichos ítems presentan DIF no uniforme. Además, en los ítems 3, 7, 9, 10 y 13, el modelo de no DIF no se ajusta bien a los datos, indicando que la respuesta a dichos ítems no sólo depende del nivel de habilidades sociales sino también del género.

Para el factor IV, se detectó DIF en siete de los trece ítems (ver tabla 7). En dichos ítems el componente de DIF fue estadísticamente significativo, siendo necesario para explicar la distribución de los datos en la tabla de contingencia.

Por último, la tabla 8 muestra los resultados obtenidos para el factor V. En este factor los ítems que presentaron evidencia de DIF fueron los ítems número 2, 5, 6, 7 y 11. En estos ítems, el modelo de no DIF no se ajusta bien a los datos ($p<0.05$), siendo el modelo de DIF más apropiado.

Acuerdo entre los dos procedimientos. El acuerdo entre los dos procedimientos fue alto. El número de ítems que fueron detectados con DIF por los dos procedimientos empleados (*MH* y modelos Logit) fue de 24 y el número de ítems detectados como no DIF fue de 60. El porcentaje de acuerdo entre los dos procedimientos fue bastante alto (95.45%), indicando que ambos procedimientos detectaron la presencia de DIF en los mismos ítems. Además el procedimiento basado en los modelos Logit detectó 4 ítems con DIF, que el estadístico de *MH*, dada su limitación para evaluar DIF no uniforme, no detectó.

Conclusiones

A pesar de que en la actualidad existe un gran número de procedimientos cada vez más sofisticados para detectar el DIF en sus diversas formas

(Fidalgo, 1996; Gómez e Hidalgo, 1997; Hidalgo y Gómez, 1999; Millsap y Everson, 1993; Potenza y Dorans, 1995), es preciso tener en cuenta que estas técnicas sólo son apropiadas para detectar el sesgo potencial en un ítem y no ofrecen por sí mismas una explicación de las causas del DIF (Camilli, 1993; Camilli y Shepard, 1994; Donoghue y Allen, 1993; Mellenbergh, 1989). Por ello, una correcta interpretación del DIF debe ser el resultado de una conjunción entre los procedimientos estadísticos y las revisiones teóricas o juicios de expertos (Shepard, 1981; Tittle, 1982). Para afirmar que un ítem está sesgado contra un determinado grupo es necesario hacer referencia a las causas o razones por las que funciona de forma distinta en dicho grupo y si esas diferencias son o no parte legítima del constructo objeto de medición. Esto implica examinar el contenido de los ítems en los cuales se encontró evidencia de DIF, e intentar dar una explicación del mismo.

En general, los resultados encontrados, en el análisis del impacto, reflejan diferencias entre varones y mujeres en determinadas habilidades sociales, principalmente en aquellas situaciones donde el adolescente se relaciona con personas de la misma edad del mismo o distinto género. El análisis del DIF permite establecer si esas diferencias encontradas son parte legítima del constructo que estamos tratando medir o por el contrario se debe a otras variables que no son relevantes en el mismo.

Para los ítems que presentan DIF en el factor I el contenido de los mismos se refiere a situaciones de interacción del adolescente con camareros o dependientes. Para los ítems que muestran DIF en el factor II, el contenido de los mismos hace referencia a la capacidad del adolescente para pedir opiniones al padre y para pedir favores o hacer cumplidos a la madre. En el factor III los ítems que presentan DIF expresan, en general, la habilidad del adolescente para relacionarse con personas de la misma edad pero de género contrario. Dada la propia formulación del ítem, donde independientemente del género del adolescente se le pregunta por su dificultad para expresar sus derechos a un chico o una chica, cabría esperar un comportamiento distinto en chicos y chicas. Por ejemplo, cuando la interacción se refiera a una chica, la respuesta que de una chica irá referida a su dificultad para expresar sus derechos a una persona del mismo género; por el contrario, en el caso de que sea un chico el que responda a dicho ítem, esta dificultad se expresará con respecto a una persona de género contrario. Estos ítems deberían reformularse, a fin de considerar las dificultades propias que puede tener un adolescente para relacionarse con personas de género contrario. Los ítems de los factores IV y V en los que se detectó la presencia de DIF presentan, en general, un contenido similar al de los ítems detectados con DIF en el factor III, reflejando la misma situación de interacción social que ya hemos comentado.

Los resultados obtenidos ponen de manifiesto que, independientemente de que puedan existir diferencias reales en habilidades sociales entre chicos y chicas de 16 a 18 años, antes de poder hacer las comparaciones oportunas y extraer algunas conclusiones al respecto, es necesario proceder a una purificación de la escala y aplicación de nuevo a una muestra de adolescentes. El proceso de depuración de la prueba supone revisar y/o eliminar los ítems que muestran un funcionamiento diferencial en los distintos grupos comparados. Apuntando la necesidad de realizar estudios de DIF en el propio proceso de construcción de tests.

Referencias

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

Bishop, Y.M.M., Fienberg, S.E. y Holland, P.W. (1975). *Discrete multivariate analysis:Theory and practice.* Cambridge, MA: MIT Press.

Caballo, V.E. (1993). *Manual de evaluación y entrenamiento de las habilidades sociales.* Madrid: Siglo XXI.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues?. En P. N. Holland y H. Wainer (Eds), *Differential item functioning* (pp. 397-413). Hillsdale, NJ: LEA.

Camilli, G. y Shepard, L.A. (1994). *Methods for identifying biased test items*. Newbury Park, C.A.: Sage.

De Maris, A. (1991). A framework for the interpretation of first-order interaction in logit modeling. *Psychological Bulletin, 110*, 557-570.

De Maris, A. (1992). *Logit Modeling: Practical Applications*. Beverly Hill, CA: Sage.

Donoughe, J.R. y Allen, N.L. (1993). Thin versus thick matching in the Mantel-Haenszel Procedure for detecting DIF. *Journal of Educational Statistics, 18(2),* 131-154.

Fidalgo, A.M. (1996). Funcionamiento diferencial de los ítems. En J. Muñiz (Coord), *Psicometría*. Madrid: Universitas.

Francis, B., Green, M. y Payne, C. (1993). *The Glim System. Release 4,* Oxford, UK: Clarendon Press

Gómez, J. e Hidalgo, M.D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: Una revisión metodológica. *Anuario de Psicología, 74,* 3-32.

Hidalgo, M.D. (1995). *Evaluación del funcionamiento diferencial del ítem en ítems dicotómicos y politómicos: Un estudio comparativo*. Universidad de Murcia: Tesis Doctoral no publicada.

Hidalgo, M.D. y Gómez, J. (1999). Técnicas de detección de funcionamiento diferencial en ítems politómicos. *Metodología de las Ciencias del Comportamiento*. (En revisión).

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

Ironson, G.H. (1982). Use of chi-square and latent-trait approaches for detecting item bias. En R.A. Berk (Ed), *Handbook of methods for detecting item bias*. Baltimore, MD: Johns Hopkins University Press.

Linn, R.L. y Harnisch, D.L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18,* 109-118.

López Pina, J.A., Hidalgo, M.D. y Sánchez, J. (1993). *Error tipo I de las Pruebas c^2 en el estudio del sesgo de los ítems.* Comunicación presentada al III Symposium de Metodología de las Ciencias del Comportamiento, Santiago de Compostela.

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

Mazor, K.M., Clauser, B.E. y Hambleton, R.K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52,* 443-451.

Mellenbergh, G.J. (1982). Contingency table models for assesing item bias. *Journal of Educational Statistics, 7,* 105-118.

Mellenbergh, G.J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127-143.

Méndez, F.X., Inglés, C., Hidalgo, M.D. y Martínez, J.A. (1998). Interpersonal Difficulty Scale for Adolescents (IDSA). Manuscrito no publicado.

Méndez, F.X., Martínez, J.A., Sánchez, S. e Hidalgo, M.D. (1995). *Escala de Habilidades Sociales para Adolescentes* (EHSPA). Universidad de Murcia.

Millsap, R.E. y Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 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,* 23-37.

Rogers, H.J. y Swaminathan, H. (1993). A comparison of Logistic Regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105-116.

Shepard, L.A. (1981). Identifying bias in test items. *New directions for testing and measurement, 11,* 79-104.

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

Tittle, C.K. (1982). Use of judgmental methods in item bias studies. En R. A. Berk (Ed). *Handbook of methods of detecting item bias* (pp. 31-63). Baltimore, MD: Johns Hopkins University Press.

Uttaro, T. y Millsap, R.E. (1994). Factors influencing the Mantel-Haenszel Procedure in the Detection of Diferential Item Functioning. *Applied Psychological Measurement, 18,* 15-25.

(Artículo recibido: 11-9-98, *aceptado*: 2-2-99)