

Análisis de la variable sexo en la escala de búsqueda de sensaciones (sss-v) empleando técnicas de funcionamiento diferencial de los ítems

Analysis of the Gender Variable in the Sensation Seeking Scales (sss-v) Using Differential Item Functioning Techniques

Análise da Variável Sexo na Escala de Busca de Sensações (sss-v) Empregando Técnicas de Funcionamento Diferencial dos Itens

Sergio Escorial Martín*

Universidad Complutense de Madrid

Doi: <http://dx.doi.org/10.12804/revistas.urosario.edu.co/apl/a.3753>

Resumen

Estudios en el área de la personalidad han encontrado que los hombres y las mujeres difieren sistemáticamente en el grado en que buscan o necesitan experimentar nuevas sensaciones. El objetivo de este estudio es analizar estas diferencias en el instrumento sss-v utilizando técnicas de funcionamiento diferencial de los ítems (DIF), con el fin de determinar si estas diferencias son el resultado de un funcionamiento anómalo de los ítems en el grupo de hombres y mujeres o si, por el contrario, están reflejando diferencias reales. Para ello, se administró el test sss-v a 1191 participantes de un amplio rango de edad. Se emplearon los siguientes métodos para detectar el posible DIF: SIBTEST, regresión logística, estadístico χ^2 de Lord y modelo DFIT. Los resultados obtenidos sugieren que, pese a existir algunos ítems con DIF, las diferencias observadas parecen

deberse realmente a una diferencia en el constructo de personalidad y no a la presencia de ítems con DIF.

Palabras clave: sss-v, diferencias de sexo, FDI, estandarización, regresión logística, Chi-cuadrado de Lord, Modelo DFIT, SIBTEST.

Abstract

Studies in the field of personality have found that men and women differ systematically in their level of sensation seeking. The aim of this study is to analyze these differences in the sss-v using differential item functioning (DIF) techniques to determine whether these differences are the result of a differential functioning of the items between males and females or if, on the contrary, they may be reflecting true differences in the assessed dimensions. To this end, 1191 participants within a wide age range were evaluated using the SSS-V test.

* Departamento de Metodología de las Ciencias del Comportamiento, Facultad de Psicología, Universidad Complutense de Madrid. Correspondencia: Departamento de Metodología de las Ciencias del Comportamiento, Facultad de Psicología, Universidad Complutense de Madrid. Campus de Somosaguas. Pozuelo de Alarcón (28223) Teléfono: 91 394 3080. Correo electrónico: sergio.escorial@psi.ucm.es

Cómo citar este artículo: Escorial-Martín, S. (2017). Análisis de la variable sexo en la escala de búsqueda de sensaciones (sss-v) empleando técnicas de funcionamiento diferencial de los ítems. *Avances en Psicología Latinoamericana*, 35(2), 387-405. doi: <http://dx.doi.org/10.12804/revistas.urosario.edu.co/apl/a.3753>

The following detection methods were used in order to examine DIF: SIBTEST, logistic regression, Lord's χ^2 test, and the DFIT model. Despite the fact that some items with DIF exist, the sex 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.

Keywords: sss-v, gender differences, DIF, standardization, logistic regression, Lord's χ^2 test, DFIT framework, SIBTEST.

Resumo

Estudos na área da personalidade têm encontrado que os homens e as mulheres diferem sistematicamente no grau em que buscam ou precisam experimentar novas sensações. O objetivo deste estudo é analisar estas diferenças no instrumento sss-v utilizando técnicas de Funcionamento Diferencial dos Itens (DIF), com o fim de determinar se estas diferenças são o resultado de um funcionamento anômalo dos itens no grupo de homens y mulheres ou se, pelo contrário, estão refletindo diferenças reais. Para isto, administrou-se o teste sss-v a 1191 participantes de um amplo rango de idade. Empregaram-se os seguintes métodos para detectar o possível DIF: SIBTEST, regressão logística, estatístico χ^2 de Lord e modelo DFIT. Os resultados obtidos sugerem que, pese a existir alguns itens com DIF, as diferenças observadas parecem dever-se realmente a uma diferença no constructo de personalidade e não à presença de itens com DIF.

Palavras-chave: sss-v, diferenças de sexo, FDI, standardização, regressão logística, χ^2 -quadrado de Lord, Modelo DFIT, SIBTEST

La literatura científica es rica en estudios dirigidos a investigar las diferencias de sexo en los rasgos de personalidad (Colom & Jayme-Zaro, 2004; Feingold, 1994; Jorm, 1987). Una de las dimensiones de personalidad considerada más frecuentemente es la búsqueda de sensaciones, constructo desarrollado por Marvin Zuckerman

y otros autores (Zuckerman, 1979, 1984, 1990; Zuckerman, Eysenck & Eysenck, 1978) y que ha merecido una considerable atención desde su desarrollo original (Chico, 2000). Zuckerman (1994) define la Búsqueda de Sensaciones como: “un rasgo definido por la búsqueda de sensaciones y emociones variadas, nuevas, complejas e intensas, y la disposición a asumir riesgos físicos, sociales, legales y financieros con el fin de obtener tales experiencias y sensaciones” (p. 27).

En la actualidad, la explicación de la dimensión de búsqueda de sensaciones se basa en un modelo influenciado por factores genéticos, bioquímicos, psicofisiológicos y sociales que influyen en ciertos comportamientos, en concreto, varios tipos de conductas de riesgo (Roberti, 2004; Roth & Hammelstein, 2003; Zuckerman, 1994, 2005). De hecho, este rasgo de personalidad se ha asociado sistemáticamente con deportes de riesgo, profesiones peligrosas, consumo de drogas, conducta sexual arriesgada o preferencias estéticas (Aluja, 1991; Chico, 2000; Gomá, 1995; Trimpop, Kerr & Kirkcaldy, 1999).

El instrumento más frecuentemente empleado para evaluar la búsqueda de sensaciones es el *Seeking Sensation Scale-Form V* (sss-v), desarrollado originalmente por Zuckerman, Eysenck y Eysenck (1978), un cuestionario de autoinforme compuesto por cuatro subescalas, cada una de ellas medida a partir de 10 ítems dicotómicos: la búsqueda de emociones y aventuras (TAS), la búsqueda de experiencia (ES), la desinhibición (DES) y la susceptibilidad al aburrimiento (BS). Las cuatro escalas contribuyen al cómputo de una puntuación total de búsqueda de sensaciones (sss).

Hasta el momento, tres importantes estudios han examinado las diferencias de sexo en búsqueda de sensaciones, empleando muestras con un amplio rango de edad. Los primeros datos globales en ss provienen de una muestra inglesa analizada por Zuckerman et al. (1978). Estos autores encontraron que la puntuación total en la escala descendía gradualmente desde el grupo de edad más joven

(16-19 años) hasta el grupo de más edad (mayores de 60 años), en ambos sexos. Además, las diferencias de sexo fueron estadísticamente significativas para todos los niveles de edad, siendo los hombres quienes puntuaban sistemáticamente más alto que las mujeres. Estos resultados iniciales fueron parcialmente apoyados por Ball, Farnhill, y Wangeman (1984), al emplear una muestra australiana con edades comprendidas entre los 17 y los 70 años. En la escala total, Ball et al. (1984) también encontraron un descenso de la puntuación, según se incrementa la edad para ambos sexos. Sin embargo, encontraron un efecto significativo de la interacción sexo y edad, con una disminución más pronunciada de la búsqueda de sensaciones en los hombres de 30-39 años, lo que dio como resultado una media ligeramente superior para las mujeres en ese mismo grupo de edad. Ball et al. (1984) especularon que las diferencias en SSS son generacionales y no relacionadas con la edad, porque las mujeres de 30-39 años “fueron criadas en condiciones de posguerra y en tiempos de dificultades económicas” (p. 264). Finalmente, los resultados de un estudio con muestra alemana realizado por Beauducel, Strobel y Brocke (2003) parece apuntar en una dirección similar. Este estudio replicó ampliamente el patrón de resultados encontrados originariamente por Zuckerman et al. (1978), con la única excepción de la subescala ES, los hombres puntuaron significativamente por encima de las mujeres en las otras tres subescalas y en el conjunto de la escala.

Empleando muestras españolas, hay que destacar el estudio de Chico (2000), si bien con un rango de edad sensiblemente menor al de los estudios anteriores, siendo su grupo de mayor edad el comprendido entre los 30 y los 39 años. Este estudio reproduce, en gran medida, los resultados anteriores. Se encontraron diferencias significativas a favor de los varones en tres de las cuatro subescalas (TAS, DES y BS) y en la puntuación total (SSS). Tanto los hombres como las mujeres presentaron con la edad un progresivo descenso de sus puntuaciones en SSS.

Todos los estudios anteriores utilizaron la escala SSS-v para medir la búsqueda de sensaciones, que puede ser considerada como una medida estándar de este rasgo (Ferrando & Chico, 2001).

Las principales propiedades psicométricas de la SSS-v se pueden encontrar en los cientos de estudios realizados en los últimos años (una síntesis de la evidencia disponible al respecto puede consultarse en Gray & Wilson, 2007 y Zuckerman, 1979, 1994, 2007a, 2007b). La validez convergente y discriminante ha sido puesta a prueba en decenas de estudios donde los resultados mostraron el grado de relación con distintos criterios (Zuckerman 2007a). La estructura factorial de la prueba SSS-v se ha replicado en muchos estudios, incluso en los que se realizan en las versiones adaptadas y traducidas a otras lenguas y culturas (Zuckerman, 2005, 2007a, 2007b). En suma, la evidencia empírica disponible parece indicar que la SSS-v es una prueba con unas propiedades psicométricas lo suficientemente buenas como para haberse convertido en una medida ampliamente utilizada en el ámbito de la personalidad.

En cuanto a los usos que se le han dado a SSS-v, cabe destacar uno en el ámbito de la selección de personal, pues aunque el cociente intelectual es considerado habitualmente como el mejor predictor del rendimiento futuro de los individuos en numerosos contextos socialmente relevantes (Jensen, 1998), en los últimos años se ha incrementado el uso de instrumentos estandarizados para evaluar dimensiones de personalidad (Corr & Mathews, 2009). Así, por ejemplo, en algunos contextos de selección de personal —como la selección de controladores de tráfico aéreo (CTA)— se utilizan puntuaciones elevadas en la dimensión de búsqueda de sensaciones como criterio para rechazar individuos (Arend, Botella, Contreras, Hernández & Santacreu, 2003; Delgado, 1995).

Dado que los tests son frecuentemente utilizados para tomar decisiones importantes en la vida de las personas, resulta esencial determinar si las diferencias observadas entre hombres y mujeres

en las puntuaciones de los test son el resultado de una diferencia real en personalidad, es decir, si son el reflejo de una diversidad que existe realmente; o si, por el contrario, estas diferencias son un mero artefacto producido por los propios ítems que componen el test, diferencias que pueden llevar a la discriminación cuando hombres y mujeres intentan acceder al mercado laboral o en algún contexto educativo. Los análisis del (DIF) pueden ayudar a examinar esta cuestión.

Los métodos estadísticos del (DIF) constituyen 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 DIF permite distinguir entre diferencias reales (impacto) y espúreas (DIF) 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 y estas diferencias reflejan diferencias reales entre los grupos en la o las variable/s medida/s por el test (Abad, Olea, Ponsoda & García, 2011; Ackerman, 1992; Fidalgo, 1996). Un ítem presenta DIF 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 este pertenezca. Esto es, en un ítem con DIF 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 personas con el mismo nivel estimado en el rasgo, es sistemáticamente mayor en un grupo que en otro durante el continuo del rasgo, se habla de DIF uniforme; en caso contrario, de DIF no uniforme.

La mayoría de los estudios llevados a cabo en el área de la personalidad y dirigidos a analizar

las diferencias de sexo se han limitado a la evaluación del impacto. Los estudios sobre la posible existencia de DIF en instrumentos comúnmente empleados para la evaluación de la personalidad son relativamente recientes y escasos (Borsboom, Mellenbergh & Van Heerden, 2002; Collins, Raju & Edwards, 2000; Ellis & Mead, 2000; Escorial & Navas, 2006, 2007; Gelin & Zumbo, 2003; Lange, Irwin, & Houran, 2000; Leo, Van Dam, Hobkirk & Earleywine, 2011; Reise, Smith, & Furr, 2001; Smith, 2002).

El objetivo principal de esta investigación es analizar las diferencias de sexo en la dimensión de búsqueda de sensaciones empleando técnicas de DIF. Para ello, en primer lugar, se procederá a analizar las diferencias de sexo en las puntuaciones observadas, seguidamente se analizará el posible DIF de los ítems de la prueba SSS-V con distintas técnicas de detección. Finalmente, se evaluará el impacto, es decir, se analizará la posible existencia de diferencias reales en el constructo analizado. Pese a ser una medida muy utilizada por los investigadores del ámbito de la personalidad, nunca antes se ha analizado el DIF en este instrumento. Lo que se persigue, básicamente, es encontrar evidencia empírica para determinar si las diferencias observadas entre varones y mujeres 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 la prueba.

Método

Participantes

La muestra estaba formada por 1191 participantes pertenecientes a la comunidad de Madrid. Para la selección de los participantes se utilizó un procedimiento incidental con sistema de cuotas para el sexo y la edad. Los datos fueron recogidos por 75 evaluadores, todos ellos licenciados universitarios convenientemente entrenados, quienes administraron el test a un hombre y una mujer en cada uno de los siguientes subgrupos de edad: <20,

20-29, 30-39, 40-49, 50-59 y >60. La media de edad para la muestra total fue 34,24 (DT = 17,27, rango 14-88). Dentro del grupo de hombres la media de edad fue 35,61 (DT = 17,54, rango 14-85) y para las mujeres fue de 32,94 (DT = 16,91, rango 16-88).

Los grupos elegidos coinciden con los utilizados en otros estudios sobre el sss-v (Ball et al., 1984; Chico, 2000; Zuckerman et al., 1978; Zuckerman & Neeb, 1980). El trabajo de campo tuvo una duración de tres semanas y la tasa de no-respuesta (personas que rehusaron participar) fue del 12%. La distribución por sexo en la muestra final fue del 48,9% para los hombres y del 51,1% para las mujeres, con un porcentaje similar para cada uno de los seis grupos de edad.

Instrumentos

Se utilizó la versión española del sss-v (Pérez & Torrubia, 1986), un test compuesto por cuatro subescalas de 10 ítems con un formato de respuesta dicotómico. La búsqueda de emociones y aventuras (TAS) es una subescala que refleja el deseo de participar en la búsqueda de emociones y actividades recreativas, arriesgadas y aventureras. La búsqueda de experiencia (ES) es una subescala que representa la necesidad de buscar nuevas experiencias por medio de, por ejemplo, viajes, drogas, música, arte o un estilo de vida poco convencional. La desinhibición (DES) es una subescala que refleja un deseo de liberación social por medio de la bebida, las fiestas o una variedad de

experiencias sexuales; mientras que los ítems de la subescala susceptibilidad al aburrimiento (BS) están apresando la aversión de los sujetos por las experiencias repetitivas.

La tabla 1 muestra los valores obtenidos en este estudio para el coeficiente Alfa de Cronbach para el índice global de búsqueda de sensaciones (sss) y para las cuatro subescalas, divididos en función del sexo, junto con los intervalos de confianza al 95% para los coeficientes, usando una aproximación mediante la distribución central F (Fan & Thompson, 2001).

Las estimaciones obtenidas son similares para hombres y mujeres, y exceden, para el test global, el punto de corte de 0,8 generalmente aceptado en el ámbito de investigación. De hecho, la fiabilidad global para la escala sss-v fue de 0,855, un valor superior al encontrado en otros estudios (véase Deditius-Island & Caruso, 2002). Respecto de las subescalas, TAS muestra el valor más alto (0,86), seguido por DES (0,75), ambos por encima del valor promedio de otros estudios, mientras que ES y BS mostraron valores inferiores (0,51 y 0,52) a los que se encontraron en estudios similares (Deditius-Island & Caruso, 2002). En cualquier caso, la fiabilidad global es aceptable, máxime cuando el propósito del análisis del DIF llevará un enfoque en el test global.

Análisis

En primer lugar, se examinaron las diferencias de sexo en las puntuaciones observadas. Para las

Tabla 1
Estimaciones de la fiabilidad para la escala sss-v de Zuckerman

	SSS		TAS		ES		DES		BS	
	h	M	h	M	h	M	h	M	h	M
α	0,87	0,84	0,86	0,86	0,65	0,57	0,76	0,72	0,53	0,51
95% IC	0,85- 0,88	0,82- 0,86	0,84- 0,87	0,84- 0,88	0,60- 0,69	0,52- 0,62	0,73- 0,79	0,69- 0,76	0,47- 0,59	0,45- 0,57

Nota: IC = intervalo de confianza

respuestas a los ítems, se empleó el estadístico χ^2 . Para la escala global, una prueba estadística de diferencia de medias para dos muestras independientes. Estas pruebas de significación estadísticas fueron acompañadas con la correspondiente medida del tamaño del efecto (d de Cohen).

En segundo lugar, se aplicaron una serie de técnicas para detectar la presencia de DIF en los ítems del SSS-V. Las técnicas utilizadas fueron la regresión logística, la prueba χ^2 de Lord, el modelo DFIT y el procedimiento SIBTEST. Dado que la prueba χ^2 de Lord y los estadísticos basados en el modelo DFIT operan dentro del marco de la teoría de respuesta al ítem (IRT), es imprescindible verificar que los datos obtenidos tienen un ajuste razonable a algún modelo de IRT. Con este propósito, se utilizó la aproximación propuesta por Hambleton y Swaminathan (1985). La estimación de parámetros se realizó con el programa MULTLOG (Thissen, 1991).

Regresión logística. Cuando la regresión logística es utilizada para estudiar el DIF (Swaminathan & Rogers, 1990), el objetivo es determinar si es suficiente introducir el nivel en el rasgo de los participantes en la función matemática para predecir la respuesta a un ítem (modelo sin DIF) o si, por el contrario, es necesario incluir un término que se refiera al grupo de pertenencia de los participantes (modelo de DIF uniforme) o un término que se refiera a la interacción entre el grupo de pertenencia de los participantes y su nivel de rasgo (modelo de DIF no uniforme).

Zumbo y Thomas (1997) proponen combinar la significación estadística con una medida del tamaño del efecto para concluir si un ítem presenta o no DIF, utilizando una estrategia de comparación de modelos. Se considera que existe algún efecto cuando el incremento en el coeficiente de pseudodeterminación (en el presente estudio se ha utilizado el de Nagelkerke) en los modelos comparados es de al menos 0,035, de modo que valores entre 0,035 y 0,070 denotarían un DIF moderado

y por encima de 0,070 un DIF severo (Jodoin & Gierl, 2001).

Estadístico χ^2 de Lord. Este estadístico está basado en la IRT y 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 DIF (Lord, 1980). Su valor ha sido determinado con el programa LINKDIF (Waller, 1998).

Modelo DFIT. Fue propuesto por Raju, van der Linden y Fleer (1995) dentro del marco de la IRT. El modelo DFIT permite estudiar no solo el DIF, sino también el funcionamiento diferencial del test completo. En concreto, este modelo ofrece un estadístico para el funcionamiento diferencial del test (DTF) y dos estadísticos de DIF: un índice compensatorio (CDIF) y otro no compensatorio (NCDIF). Se dispone de tests de significación estadística para los índices DTF y NCDIF, pero no existen para el índice CDIF. En este caso, si el índice global no fuera estadísticamente significativo, se asumirá que ninguno de los valores de CDIF lo es. Pero si el índice DTF resulta estadísticamente significativo, se eliminará el ítem con un valor absoluto para el CDIF más elevado, repitiendo este proceso ítem por ítem hasta que el índice DTF deje de ser significativo. Al alcanzar este punto, se considera que los ítems eliminados podrían presentar problemas de DIF.

Es también esencial tener en cuenta la significación práctica a la hora de interpretar los estadísticos DFIT, con independencia de que se disponga o no de una prueba de significación estadística. Se considerará que un ítem muestra DIF cuando los anteriores procedimientos iterativos lo identifiquen como tal (en el caso de CDIF), o cuando así se concluya por medio de las pruebas de significación (en el caso del NCDIF), pero solo cuando los valores obtenidos para los estadísticos (CDIF o NCDIF) superen cierto valor de corte. Los valores críticos fueron obtenidos mediante un estudio de simulación, siendo de 0,0276 para el CDIF y 0,0131

para el NCDIF. Para este procedimiento se empleó el programa LINKDIF.

SIBTEST. Este procedimiento detecta el funcionamiento diferencial de los ítems individualmente y de un conjunto de ítems simultáneamente (Shealy & Stout, 1993). Así, se comparan los promedios para los grupos focal y de referencia en un subtest compuesto por un conjunto de ítems sospechosos de mostrar DIF, lo que empareja a los participantes conforme a la puntuación obtenida en otro subtest compuesto por el resto de los ítems que, se considera, no presentan DIF (subtest válido). Para decidir qué ítems compondrían los subtest válidos y bajo sospecha se utilizó la información obtenida mediante los procedimientos de detección de DIF descritos previamente. Los análisis se realizaron con el programa Dimensionality-Based DIF/DBF Package (Stout & Roussos, 1999).

Finalmente, una vez eliminados los ítems con DIF de la prueba, se procedió a determinar si existen o no diferencias entre hombres y mujeres en la escala de búsqueda de sensaciones. Para ello, se utilizó una prueba de diferencia de medias para dos muestras independientes (acompañada con la medida del tamaño del efecto).

Resultados

Antes de realizar los análisis de DIF, se analizó la dimensionalidad del test. Este es un requisito necesario para los métodos de detección del DIF basados en la IRT (Borsboom, Mellenbergh & Van Heerden, 2002; Stout & Roussos, 1999; Clauser & Mazor, 1998; Holland & Wainer 1993). En este caso, esto implicaba poner a prueba la estructura unidimensional del instrumento sss-v, tal y como sugieren otros autores (Aluja, García & García, 2004). Para ello, se realizó un análisis factorial no lineal con el programa NOHARM (Fraser & McDonald, 1988).

La evaluación de la dimensionalidad en NOHARM se basa en la inspección visual de las covarianzas residuales tras la extracción del número prescrito

de factores y, como índice, suele utilizarse la raíz cuadrática media residual (RMSR). Tal como sugieren McDonald y Mok (1995), para este estudio se ha utilizado también la versión para mínimos cuadrados del índice normado gamma (g) de bondad de ajuste propuesto por Tanaka y Huba (1985). Los resultados obtenidos para el modelo de un factor formado por todos los ítems del sss-v (40 ítems) fueron los siguientes: la RMSR fue de 0,130 y g de 0,860. La inspección visual de las cargas factoriales de los ítems permitió identificar 12 ítems con pesos factoriales inferiores a 0,25 todos ellos (ítems 4, 8, 12, 16, 20, 28, 32, 36, 40, 18, 34 y 38). Por esta razón, se procedió a eliminar estos ítems y volver a estimar el ajuste a un modelo de una dimensión de los restantes 28 ítems del sss-v. Los resultados del ajuste obtenido para este modelo fueron 0,014, la RMSR y 0,960 el índice g . Así pues, se produce una mejora razonable en la estructura unidimensional de la prueba al pasar de 40 a 28 ítems.

Los autores entienden que resulta más apropiado trabajar con esta definición unidimensional de la escala que, aun prescindiendo de un porcentaje importante de sus ítems, opera todavía con más de 20 ítems por escala, lo que proporciona, de esta forma, un número de ítems adecuado para obtener una estimación razonablemente precisa de la característica medida por la prueba, algo muy importante para poder igualar posteriormente de forma efectiva a los sujetos de los grupos focal y de referencia y estudiar, en esos grupos igualados, la probabilidad de responder en una determinada dirección cada uno de los ítems. Todos los análisis que se realizaron a partir de este punto se condujeron para la versión unidimensional de la prueba, es decir, para la versión de 28 ítems del sss-v.

Por otra parte, para valorar el ajuste de los datos a un modelo TRI se calculó el estadístico de razón de verosimilitudes G^2 , que se distribuye como una χ^2_{N-h-1} , donde N es el número de sujetos de la muestra y h el número de parámetros a estimar. Se estimaron los parámetros de los ítems de tres modelos ITR: los modelos logísticos de 1, 2 y

3 parámetros, respectivamente. Para determinar a qué modelo en particular se ajustaban mejor los datos, se procedió a hallar la diferencia entre los valores obtenidos para G^2 en cada par de posibles modelos, ya que se trata siempre de modelos anidados (Andersen, 1973). Un valor significativo de la diferencia en G^2 indica que el modelo más complejo se ajusta mejor a los datos, mientras que, si dicho valor no es significativo, entonces el parámetro(s) extra del modelo más complejo no contribuye significativamente a explicar los datos y se escogerá, por tanto, el modelo más simple. La tabla 2 recoge los valores de G^2 de todos los modelos IRT estimados, así como los resultados de la comparación de estos. En la columna C se indica cuál es la comparación que se está realizando; así, por ejemplo, 1-2 denotaría que la comparación de esa fila sería la del modelo listado en primer lugar (1PLM) con el modelo listado en segundo lugar (2PLM). En la siguiente columna aparece la diferencia en el G^2 de esos modelos que se distribuye con $gl_1 - gl_2$ grados de libertad. A continuación, se recoge el valor de p para cada comparación.

Tal y como se desprende de esta tabla (columna MOD), el modelo IRT que mejor ajuste tuvo a los datos fue el modelo logístico de dos parámetros (2PLM), ya que dicho modelo se ajusta significativamente mejor que el de un parámetro (1PLM), y no presenta diferencias significativas en el ajuste con respecto al modelo más complejo (3PLM). Por esta razón, en los posteriores métodos de detección del DIF basados en la IRT, los parámetros de los ítems en el grupo de hombres y mujeres fueron estimados según este modelo (2PLM).

Tabla 2
Comparación de modelos TRI en el SSS-V

n	Modelos			Comparación de modelos			
	G^2	gl	c	dif G^2	gl	p	mod
28	1PLM: 19266,2	1162	1-2	877,2	28	< 0,001	
	2PLM: 18389,0	1134	1-3	876,9	56	< 0,001	2PLM
	3PLM: 18388,7	1106	2-3	0,3	28	0,999	

Diferencias entre grupos en puntuaciones observadas

La tabla 3 presenta los estadísticos descriptivos tanto a nivel de ítem como a nivel de la escala global para ambos grupos. La tabla presenta también información de tipo inferencial que permite valorar qué ítems muestran una relación estadísticamente significativa ($p < 0,01$) entre el sexo y la respuesta al ítem. Es decir, se puede observar qué grupo obtuvo una puntuación mayor y el tamaño del efecto. Estos resultados indican que entre los ítems analizados de la escala SSS-V existe una alta proporción de ítems con diferencias estadísticamente significativas a favor de los hombres, encontrándose diferencias significativas en la escala global (SSS) a favor de los hombres ($t(1189) = 5,485; p < 0,001; d = 0,3193$).

DIF

Para determinar si las diferencias permanecen o desaparecen, al comparar hombres y mujeres previamente emparejados en cuanto a su nivel de búsqueda de sensaciones, se llevó a cabo un análisis con objeto de detectar el posible funcionamiento diferencial. La tabla 4 muestra los resultados estadísticos obtenidos con los métodos previamente descritos al ser aplicados a los ítems de la escala SSS-V. Además, dado que algunos procedimientos se basan en la IRT, se presentan también en la tabla 4 los parámetros de los ítems estimados según el 2PLM tanto para el grupo de hombres como para el de mujeres.

El análisis de regresión logística detectó DIF uniforme en ocho ítems, pero únicamente los

Tabla 3
Estadísticos descriptivos y diferencias de sexo para la escala y los ítems

	Hombres	Mujeres	t (p)	d
N	582	609		
Alfa	0,89	0,87		
Media escala (DT)	13,62 (6,45)	11,62 (6,08)	5,485 (<0,001)	-0,3193
Asimetría escala (ET)	-0,07 (0,10)	-0,03 (0,10)		
Curtosis escala (ET)	-0,89 (0,20)	-0,86 (0,20)		
Media ítems (DT)	Hombres	Mujeres	c ² (p)	d
Ítem 1	0,32 (0,468)	0,20 (0,398)	24,634 (<0,001)	-0,2767
Ítem 2	0,39 (0,487)	0,42 (0,494)	1,409 (0,235)	0,0612
Ítem 3	0,47 (0,500)	0,44 (0,496)	1,683 (0,195)	-0,0602
Ítem 5	0,66 (0,475)	0,60 (0,491)	4,891 (0,027)	-0,1242
Ítem 6	0,73 (0,445)	0,58 (0,494)	28,483 (<0,001)	-0,3187
Ítem 7	0,20 (0,402)	0,08 (0,277)	34,614 (<0,001)	-0,3490
Ítem 9	0,61 (0,487)	0,54 (0,499)	6,817 (0,009)	-0,1419
Ítem 10	0,42 (0,493)	0,37 (0,484)	2,312 (0,128)	-0,1024
Ítem 11	0,14 (0,348)	0,11 (0,307)	3,547 (0,060)	-0,0916
Ítem 13	0,57 (0,496)	0,49 (0,500)	7,205 (0,007)	-0,1606
Ítem 14	0,21 (0,410)	0,15 (0,360)	7,274 (0,007)	-0,1557
Ítem 15	0,35 (0,479)	0,24 (0,429)	17,532 (<0,001)	-0,2422
Ítem 17	0,69 (0,464)	0,48 (0,500)	54,415 (<0,001)	-0,435
Ítem 19	0,81 (0,393)	0,66 (0,473)	32,497 (<0,001)	-0,3442
Ítem 21	0,64 (0,481)	0,52 (0,500)	17,141 (<0,001)	-0,2445
Ítem 22	0,74 (0,441)	0,71 (0,454)	1,144 (0,285)	-0,0670
Ítem 23	0,24 (0,430)	0,15 (0,362)	15,048 (<0,001)	-0,2269
Ítem 24	0,39 (0,487)	0,31 (0,461)	8,678 (0,003)	-0,1488
Ítem 25	0,56 (0,497)	0,46 (0,499)	11,976 (0,001)	-0,2008
Ítem 26	0,40 (0,489)	0,45 (0,498)	3,652 (0,056)	0,1013
Ítem 27	0,62 (0,485)	0,56 (0,497)	5,008 (0,025)	-0,1222
Ítem 29	0,37 (0,484)	0,26 (0,440)	16,189 (<0,001)	-0,2381
Ítem 30	0,40 (0,489)	0,63 (0,482)	67,870 (<0,001)	0,4738
Ítem 31	0,28 (0,448)	0,22 (0,418)	4,236 (0,040)	-0,1386
Ítem 33	0,56 (0,497)	0,48 (0,500)	6,533 (0,011)	-0,1605
Ítem 35	0,81 (0,390)	0,65 (0,476)	38,378 (<0,001)	-0,3669
Ítem 37	0,59 (0,492)	0,51 (0,500)	7,135 (0,008)	-0,1613
Ítem 39	0,46 (0,498)	0,34 (0,474)	17,060 (<0,001)	-0,2470

Tabla 4
Ítems con posible DIF detectados por diferentes procedimientos

Ítem	TRI hombres		TRI mujeres		Técnica de detección de DIF				DIF
	a	b	a	b	rl	SIBTEST* (β)	c ² Lord	ncdif	
1*	1,12	0,88	1,50	0,94	$\Delta\chi^2_{(2)} = 6,84 (0,033)$; $\Delta R^2: .006$	0,060 (0,018)	6,10 (0,047)	.0017 (<0,001)	No DIF
2	0,58	0,90	0,66	0,23	$\Delta\chi^2_{(2)} = 10,18 (0,006)$; $\Delta R^2: 0,010$	-0,097 (0,001)	25,35 (<0,001)	0,0145 (<0,001)	DIF uniforme moderado - M
3	2,18	0,15	1,90	-0,07	$\Delta\chi^2_{(2)} = 8,65 (0,013)$; $\Delta R^2: 0,006$	-0,082 (0,002)	30,69 (<0,001)	0,0193 (<0,001)	DIF uniforme moderado - M
5	1,66	-0,54	1,95	-0,58	$\Delta\chi^2_{(2)} = 2,16 (0,340)$; $\Delta R^2: 0,001$	-0,036 (0,160)	7,29 (0,026)	0,0026 (0,028)	No DIF
6*	0,80	-1,37	0,76	-0,77	$\Delta\chi^2_{(2)} = 14,66 (<0,001)$; $\Delta R^2: 0,015$	0,094 (0,001)	26,70 (<0,001)	0,0298 (<0,001)	DIF uniforme moderado - H
7*	1,38	1,38	1,47	1,74	$\Delta\chi^2_{(2)} = 12,70 (0,002)$; $\Delta R^2: 0,015$	0,064 (0,001)	3,80 (0,150)	0,0004 (<0,001)	DIF uniforme pequeño - H
9*	2,01	-0,33	2,55	-0,38	$\Delta\chi^2_{(2)} = 0,64 (0,726)$; $\Delta R^2: 0,000$	-0,016 (0,549)	11,49 (0,003)	0,0042 (0,005)	No DIF
10	1,64	0,36	1,48	0,19	$\Delta\chi^2_{(2)} = 3,26 (0,196)$; $\Delta R^2: 0,002$	-0,064 (0,016)	17,91 (<0,001)	0,0100 (<0,001)	No DIF
11	1,74	1,57	1,12	1,92	$\Delta\chi^2_{(2)} = 6,39 (0,041)$; $\Delta R^2: 0,009$	-0,021 (0,267)	3,50 (0,174)	0,0002 (<0,001)	No DIF
13*	1,93	-0,17	2,74	-0,23	$\Delta\chi^2_{(2)} = 2,15 (0,341)$; $\Delta R^2: 0,001$	-0,025 (0,350)	18,63 (<0,001)	0,0069 (0,020)	No DIF
14*	2,10	1,08	1,23	1,39	$\Delta\chi^2_{(2)} = 4,96 (0,084)$; $\Delta R^2: 0,005$	-0,025 (0,255)	6,72 (0,035)	0,0008 (<0,001)	No DIF
15*	2,64	0,51	2,08	0,56	$\Delta\chi^2_{(2)} = 0,39 (0,823)$; $\Delta R^2: 0,000$	0,007 (0,756)	3,66 (0,160)	0,0015 (<0,001)	No DIF
17*	1,23	-0,79	2,13	-0,20	$\Delta\chi^2_{(2)} = 39,08 (<0,001)$; $\Delta R^2: 0,030$	0,148 (<0,001)	86,82 (<0,001)	0,0786 (<0,001)	DIF uniforme moderado - H
19*	1,46	-1,33	1,62	-0,87	$\Delta\chi^2_{(2)} = 16,16 (<0,001)$; $\Delta R^2: 0,014$	0,087 (<0,001)	40,09 (<0,001)	0,0407 (<0,001)	DIF uniforme pequeño - H
21*	1,90	-0,43	2,19	-0,33	$\Delta\chi^2_{(2)} = 2,65 (0,266)$; $\Delta R^2: 0,002$	0,030 (0,268)	8,04 (0,018)	0,0043 (<0,001)	No DIF
22	0,99	-1,22	0,93	-1,39	$\Delta\chi^2_{(2)} = 1,90 (0,387)$; $\Delta R^2: 0,002$	-0,033 (0,203)	1,37 (0,505)	0,0007 (<0,001)	No DIF
23*	0,82	1,62	0,56	2,87	$\Delta\chi^2_{(2)} = 5,13 (0,077)$; $\Delta R^2: 0,006$	0,042 (0,081)	1,01 (0,603)	0,0008 (<0,001)	No DIF
24*	0,55	0,95	0,51	1,39	$\Delta\chi^2_{(2)} = 2,08 (0,353)$; $\Delta R^2: 0,002$	0,027 (0,345)	3,08 (0,214)	0,0019 (<0,001)	No DIF
25*	2,34	-0,12	2,87	-0,15	$\Delta\chi^2_{(2)} = 1,00 (0,607)$; $\Delta R^2: 0,001$	0,002 (0,935)	9,05 (0,011)	0,0032 (0,044)	No DIF
26	0,70	0,72	0,70	0,02	$\Delta\chi^2_{(2)} = 17,97 (<0,001)$; $\Delta R^2: 0,018$	-0,127 (<0,001)	43,74 (<0,001)	0,0267 (<0,001)	DIF uniforme moderado - M
27	0,95	-0,59	0,87	-0,61	$\Delta\chi^2_{(2)} = 1,97 (0,373)$; $\Delta R^2: 0,002$	0,001 (0,962)	0,11 (.949)	0,0001 (0,226)	No DIF
29*	1,64	0,53	1,91	0,52	$\Delta\chi^2_{(2)} = 1,42 (0,492)$; $\Delta R^2: 0,001$	0,020 (0,432)	4,54 (0,104)	0,0013 (0,448)	No DIF
30*	0,82	0,64	1,23	-0,85	$\Delta\chi^2_{(2)} = 142,60 (<0,001)$; $\Delta R^2: 0,131$	-0,336 (<0,001)	241,43 (<0,001)	0,1999 (<0,001)	DIF uniforme severo - M
31	0,63	1,70	0,62	1,81	$\Delta\chi^2_{(2)} = 0,16 (0,923)$; $\Delta R^2: 0,000$	0,004 (0,894)	0,24 (0,885)	0,0001 (<0,001)	No DIF
33	0,84	-0,28	0,94	-0,21	$\Delta\chi^2_{(2)} = 0,17 (0,919)$; $\Delta R^2: 0,000$	0,014 (0,626)	3,79 (0,151)	0,0020 (<0,001)	No DIF
35*	1,19	-1,53	1,27	-0,92	$\Delta\chi^2_{(2)} = 21,23 (<0,001)$; $\Delta R^2: 0,019$	0,110 (<0,001)	42,32 (<0,001)	0,0485 (<0,001)	DIF uniforme moderado - H
37*	2,27	-0,23	2,20	-0,31	$\Delta\chi^2_{(2)} = 1,13 (0,568)$; $\Delta R^2: .001$	-0,015 (0,558)	5,21 (0,074)	0,0034 (<0,001)	No DIF
39*	1,01	0,26	0,95	0,52	$\Delta\chi^2_{(2)} = 3,95 (0,139)$; $\Delta R^2: 0,003$	0,033 (0,250)	3,87 (0,144)	0,0025 (<0,001)	No DIF

ítems 17 y 30 mostraron un nivel de DIF cercano a moderado y severo, respectivamente. Los resultados obtenidos por el método del SIBTETS (a nivel de detección de ítems individualmente) fueron similares, mostrando nueve ítems con DIF uniforme (los mismos encontrados mediante regresión logística más el ítem 3). No obstante, los tamaños del efecto de DIF, estimados mediante la magnitud de β (Roussos & Stout, 1996), fueron sistemáticamente mayores que los obtenidos con el anterior procedimiento. Además, no se detectó la existencia de DIF no uniforme en ningún ítem.

Los resultados obtenidos al trabajar con el modelo DFIT ponen de manifiesto que no existe un funcionamiento diferencial de la escala SSS-V en su conjunto ($DTF = 1,189$; $p = 0,466$). Por tanto, los ítems con valores para el índice CDIF por encima del valor de corte no fueron tenidos en cuenta. Por su parte, al utilizar el índice NCDIF se detectaron ocho ítems con DIF, es decir, se detectó la presencia estadísticamente significativa de DIF en ocho ítems en los que, además, el valor del NCDIF superaba el punto de corte anteriormente especificado en el apartado de método. Finalmente, se detectó la presencia de DIF en diez ítems al aplicar el estadístico χ^2 de Lord.

Por otra parte, la tabla 5 presenta de forma resumida la información más relevante en relación con los análisis llevados a cabo al usar el procedimiento SIBTEST y dirigidos a analizar el efecto del DIF en un conjunto de ítems considerados simultáneamente. En ella se muestran el número de ítems en el subtest válido y los ítems del subtest

bajo sospecha. El subtest bajo sospecha estaba compuesto por aquellos ítems identificados con DIF, por medio de la regresión logística o el propio SIBTEST (ítems detectados individualmente), o por estos procedimientos más algún índice basado en la IRT. La razón de enfatizar los resultados obtenidos mediante métodos no basados en la IRT se relaciona con el hecho de que el tamaño de la muestra, aun siendo adecuado para trabajar dentro del marco de la IRT, no era excesivamente grande y el error de estimación podría estar cumpliendo un papel en los resultados. En la tabla se indican además qué ítems mostraron DIF uniforme a favor del grupo de hombres y cuáles lo hicieron a favor del grupo de mujeres y la probabilidad asociada con el estadístico de contraste que evalúa el funcionamiento diferencial del subtest bajo sospecha.

Se observa que nueve ítems mostraron DIF en la escala SSS-V; cuatro de ellos con DIF uniforme a favor del grupo de mujeres (ítems 2, 3, 26 y 30) y cinco ítems con DIF uniforme a favor del grupo de hombres (ítems 6, 7, 17, 19 y 35). En cualquier caso, no se detectó funcionamiento diferencial en el test bajo sospecha ($p = 0,196$). Esto puede deberse al efecto de cancelación, en tanto la escala cuenta con ítems que favorecen en algunos casos al grupo de mujeres y en otros casos al grupo de hombres.

Si se examinan conjuntamente las tablas 4 y 5, se hace evidente la gran consistencia de los diversos procedimientos utilizados para detectar el DIF. Por ejemplo, siete de los nueve ítems que conformaban el subtest bajo sospecha fueron detectados como ítems con DIF por todos los procedimientos

Tabla 5
Resultados obtenidos mediante el procedimiento SIBTEST

N° de ítems en el subtest válido	Ítems en el subtest bajo sospecha	dif uniforme a favor de		
		h	m	dtf (p)
19	2, 3, 6, 7, 17, 19, 26, 30, 35	6, 7, 17, 19 y 35	2, 3, 26 y 30	.196

utilizados (ítems 2, 6, 17, 19, 26, 30 y 35) y otros dos (ítems 3 y 7) fueron etiquetados con DIF por al menos tres procedimientos. Únicamente los ítems 10 y 13 fueron etiquetados con DIF por un solo procedimiento (estadístico χ^2 de Lord).

Para ilustrar el tipo de DIF detectado, así como para dar idea de su magnitud, se examinaron también las curvas características de los ítems con DIF en el grupo de varones y en el de mujeres. La figura 1 recoge estas representaciones gráficas que sirven para complementar la información estadística recogida en las tablas 4 y 5. Dicha figura permite ver gráficamente que todos los ítems detectados con DIF mostraron DIF uniforme y que cinco lo hicieron a favor del grupo de hombres (ítems 6, 7, 17, 19 y 35), mientras que cuatro lo hicieron a favor del grupo de mujeres (ítems 2, 3, 26 y 30).

Impacto

A la vista de estos resultados sería interesante dar respuesta a la pregunta de si los ítems que presentan DIF son, al menos en parte, los responsables de las diferencias encontradas entre ambos sexos al analizar la escala de manera global. Para responder a esta cuestión, se seleccionaron los ítems que figuran en la tabla 5 como ítems del subtest válido, esto es, ítems de los que se tienen razones fundamentadas para suponer que no operan diferencialmente en función del sexo y que proporcionan, en principio, una medida válida de la característica evaluada. A continuación, se calculó en dicho subtest la puntuación media de cada sexo y se realizó la correspondiente comparación de medias. La media en una escala formada por los ítems que componen el subtest válido en el grupo de los hombres fue 8,42 con una desviación típica de 4,75, mientras que en el grupo de mujeres esa media fue de 7,22 y la desviación típica fue de 4,30. Esa diferencia resultó ser estadísticamente significativa ($t(1189) = 5,706$; $p < 0,001$; $d = 0,3315$). Nótese que el tamaño del efecto es similar al que

se obtiene cuando se analiza la escala global (ver tabla 2).

Discusión

El objetivo principal de esta investigación era establecer si las diferencias de sexo observadas en la dimensión de búsqueda de sensaciones son diferencias genuinas en el constructo o si tales diferencias están causadas por un funcionamiento diferencial de los ítems, trabajando con un instrumento ampliamente utilizado pero que, hasta donde se sabe, nunca antes había sido analizado desde la perspectiva del DIF.

Tal y como se señaló anteriormente, el análisis inicial de la dimensionalidad del instrumento condujo a la eliminación de 12 ítems. Conviene resaltar que la mayoría de los ítems eliminados pertenecen a dos subescalas del instrumento SSS-V: las subescalas de ES y BS. La fiabilidad de esas subescalas en los datos de esta investigación es relativamente baja. Este resultado es congruente con el extenso estudio realizado por Deditius-Island y Caruso (2002), quienes analizaron 244 investigaciones empíricas que empleaban el SSS-V, donde se encontró que sistemáticamente las subescalas de ES y BS mostraban los menores coeficientes de fiabilidad.

En relación con este aspecto, es importante subrayar que el hecho de que la mayoría de los ítems eliminados pertenezcan a las subescalas mencionadas podría suponer una amenaza a la validez de contenido del constructo de búsqueda de sensaciones que se analizó en el presente estudio. No obstante, conviene recordar que, (i) los doce ítems fueron eliminados porque presentaban cargas factoriales inferiores a 0,25 en el análisis factorial confirmatorio realizado con NOHARM, (ii) los 28 ítems del SSS-V analizados en este estudio mostraban un buen ajuste a un modelo formado por una única dimensión muy robusta, y (iii) el índice de correlación entre la puntuación global de

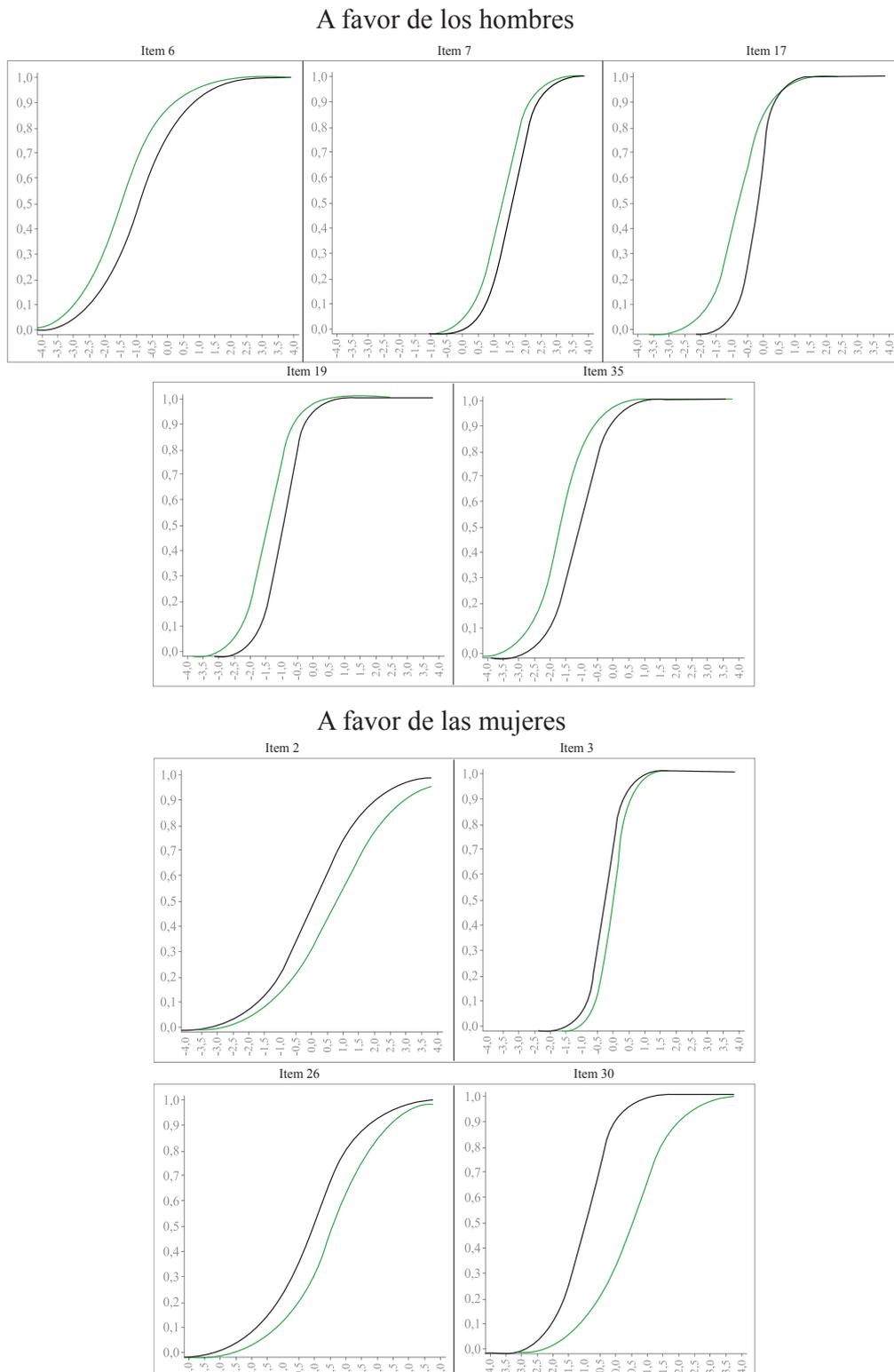


Figura .ICCI de los ítems detectados con DIF para ambos sexos

búsqueda de sensaciones de la versión completa del sss-v (40 ítems) y la puntuación global en la versión de los 28 ítems analizados en este estudio fue de 0,969.

Los resultados obtenidos en el análisis acerca de las diferencias de sexo en las puntuaciones observadas mostraron que se encontraron diferencias significativas en el índice global de búsqueda de sensaciones (sss) y en un número bastante elevado de los ítems que componen el sss-v. La dirección de tales diferencias fue consistente con la literatura científica previa: los hombres tuvieron una mayor puntuación que las mujeres en búsqueda de sensaciones (Zuckerman et al., 1978; Ball et al., 1984; Beaducel et al., 2003).

El estudio del funcionamiento diferencial de los ítems mostró que las diferencias de sexo encontradas inicialmente en el análisis de los ítems no fueron consecuencia de problemas derivados del instrumento de medición. En efecto, se detectó la presencia de DIF en un número reducido de ítems, en todos ellos el DIF fue uniforme y el tamaño del efecto fue, por otra parte, moderado en la mayoría de los casos. Además, se podría estar observando un efecto de cancelación, ya que cinco ítems presentan DIF uniforme a favor del grupo de hombres, mientras que cuatro presentan DIF uniforme a favor del grupo de mujeres. Este efecto de cancelación se traduce en que la escala en su conjunto no funciona de forma diferente en el grupo de hombres y mujeres.

La mayoría de los ítems en los que se ha detectado DIF han sido identificados consistentemente (con al menos cuatro de los cinco métodos de detección) ocho ítems de los nueve que forman parte del subtest sospechoso (2, 3, 6, 17, 19, 26, 30 y 35).

Para completar los resultados obtenidos, se llevó a cabo un análisis dirigido a responder a la pregunta de si los ítems que funcionan diferencialmente eran, al menos en parte, responsables de las diferencias encontradas entre varones y mujeres, al analizar la escala globalmente. Para

dar respuesta a esta cuestión, tal y como se recoge en el apartado de resultados, se seleccionaron los ítems del subtest válido empleado en el SIBTEST, esto es, los ítems de los que se tienen fundadas razones para suponer que no operan diferencialmente en función del sexo 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 y se realizó la correspondiente comparación de medias.

Es decir, con este análisis se trató de comprobar si una vez eliminados los ítems con DIF del test, existían diferencias entre hombres y mujeres en la variable medida. En caso de existir esas diferencias, estas se podrían denominar en este caso como ‘impacto’. Los resultados mostraron que la tendencia no cambia cuando los ítems que presentan DIF fueron eliminados. En otras palabras, las diferencias de sexo que se encontraban inicialmente se siguen manteniendo. Más aun, los tamaños del efecto encontrados para tales diferencias fueron muy similares en la escala original y en la escala formada a partir de los ítems del subtest válido.

Cuando se realiza un análisis sobre el contenido de los ítems identificados con DIF, se observan algunas características que son pertinente resaltar. Así por ejemplo, parece que hombres y mujeres (igualados en la característica evaluada) difieren en su probabilidad de responder al ítem en la dirección del rasgo cuando este incluye en su redacción a determinados colectivos marginales o de opción sexual no mayoritaria. Así, por ejemplo, se observa DIF a favor de las mujeres en los ítems 26 (“Me gustaría hacer amigos procedentes de grupos marginales”) o, de especial magnitud, el ítem 30 (“Me gustaría conocer personas homosexuales (hombres o mujeres)”). Sin embargo, se observa DIF a favor del grupo de hombres en ítems que incluyen una clara referencia al sexo o a actividades peligrosas y/o de riesgo. Como ejemplo, en los ítems 7 (“Me gusta la compañía de personas

liberadas que practican el cambio de parejas”), 17 (“Me gustaría aprender a volar en avioneta”), 19 (“Me gusta salir con personas del sexo opuesto que sean físicamente excitantes”) o 35 (“Me gusta ver las escenas sexys de las películas”).

Estos resultados podrían tener algunos correlatos sociales de cierta importancia. Por ejemplo, y congruentemente con la literatura previa, los hombres podrían ser más vulnerables al desarrollo de comportamientos antisociales, debido, entre otras muchas razones, a su mayor nivel de búsqueda de sensaciones (Aluja, 1991; Chico, 2000; Gomá, 1995; Trimpop, Kerr & Kirkcaldy, 1999). Por consiguiente, y siguiendo a Lykken (1995), el esfuerzo que se requiere para socializar correctamente a los individuos sería, en líneas generales, en promedio y como grupo, mayor para los hombres que para las mujeres.

En resumen, los resultados del presente estudio proporcionan evidencia empírica para apoyar el hecho de que las diferencias de sexo encontradas en el sss-v no son el resultado de sesgos importantes en el instrumento de medida. Estos resultados tienen implicaciones importantes para la validez de las puntuaciones en esta prueba, ya que representan un paso más en el proceso de obtención de pruebas favorables en apoyo del uso de las puntuaciones obtenidas en contextos aplicados.

No obstante, aunque según los resultados de este estudio la presencia de ítems con DIF en la sss-v no estarían incidiendo en las diferencias de sexo que se encuentran en el constructo de búsqueda de sensaciones, lo que sí pueden sugerir es la conveniencia de que en las investigaciones aplicadas se considere sistemáticamente la necesidad de utilizar baremos diferenciados para hombres y mujeres en búsqueda de sensaciones. En este sentido, conviene remarcar que, si bien, hasta donde se conoce en esta investigación, no existen baremos publicados de la sss-v, todos los estudios revisados analizan las distribuciones de las puntuaciones en el grupo de hombres y en el grupo de mujeres por separado

(Chico, 2000; Colom & Jayme-Zaro, 2004; Pérez & Torrubia, 1986; Zuckerman, 2007a).

Por otra parte, los resultados de algunos estudios relativamente recientes, como el de Stark, Chernyshenko y Drasgow (2004), ponen de manifiesto que la repercusión práctica de la presencia de un funcionamiento diferencial del test (DTF) en diferentes contextos aplicados como la selección de personal o la admisión en determinadas instituciones educativas no tiene prácticamente ningún efecto. Es decir, que aunque muchos ítems presentaran DIF y DTF, la repercusión de este hecho en las potenciales decisiones de selección sería mínima. Si esto es así con test en los que se ha detectado DIF en muchos ítems y un DTF estadísticamente significativo, el panorama en estos contextos será todavía mucho más favorable si se emplean pruebas que presenten DIF en muy pocos ítems y/o que además no presenten un DTF significativo, como según los resultados del presente estudio es el caso de la sss-v.

Finalmente, el presente estudio tiene algunas limitaciones. Quizá, la más relevante esté relacionada con la muestra de participantes analizada. En este sentido, en el presente estudio no se han considerado variables como profesión, nivel de estudios, nivel socioeconómico, consumo de drogas, estado civil, etc. Dado que en este estudio se analizan diferencias de sexo en búsqueda de sensaciones desde la perspectiva del DIF, sería conveniente contrastar que ambos grupos en función de sexo son equivalentes en estas variables, ya que podría suceder que los resultados encontrados en el presente estudio se deban en realidad a estas variables sociodemográficas que no han sido controladas. Por esta razón, futuros estudios se deberían dirigir a contrastar si se reproduce este patrón de resultados (i) incrementando sustancialmente el N, (ii) controlando la equivalencia de hombres y mujeres en estas variables sociodemográficas anteriormente mencionadas y (iii) empleando otras medidas del constructo de búsqueda de sensaciones.

Referencias

- Abad, F. J., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en Ciencias Sociales y de la Salud*. Madrid: Síntesis.
- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.
- Aluja, A. (1991). *Personalidad desinhibida, agresividad y conducta antisocial*. Barcelona: PPU.
- Aluja, A., García, O., & García, L. F. (2004). Exploring the structure of Zuckerman's sensation seeking scale, Form V in a Spanish sample. *Psychological Reports, 95*, 338-344.
- Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123-140.
- Arend, I., Botella, J., Contreras, M. J., Hernández, J. M., & Santacreu, J. (2003). A betting dice test to study the interactive style of risk-taking behavior. *Psychological Records, 53*, 217-230.
- Ball, I. L., Farnill, D., & Wangemen, J. F. (1984). Sex and age differences in sensation seeking: Some national comparisons. *British Journal of Personality, 75*, 257-265.
- Beauducel, A., Strobel, A., & Brocke, B. (2003). Psychometrische Eigenschaften und Normen einer deutschsprachigen Fassung der Sensation Seeking Skalen, Form V [Psychometric properties and norms of a German version of the Sensation Seeking Scales, Form V.], *Diagnostica, 49*, 61-72.
- Borsboom, D., Mellenbergh, G., & Van Heerden, J. (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement, 26*, 433-450.
- Chico, E. (2000). Búsqueda de Sensaciones [Sensation Seeking]. *Psicothema, 12*, 229-235.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues & Practice, 17*, 31-44.
- Collins, W., Raju, N., & Edwards, J. (2000). Assessing differential functioning in a Satisfaction scale. *Journal of Applied Psychology, 85*, 451-461.
- Colom, R., & Jayme, Z. M. J. (2004). *La psicología de las diferencias de sexo*. Madrid: Biblioteca Nueva.
- Corr, P. J., & Matthews, G. (2009). *The Cambridge Handbook of Personality Psychology*. New York: Cambridge University Press.
- Deditius-Island, H. K., & Caruso, J. C. (2002). An examination of the reliability of scores from Zuckerman's Sensation Seeking Scales, Form V. *Educational and Psychological Measurement, 62*(4), 728-734.
- Delgado, C. (1995). Sesgo de género en la medición del neuroticismo [Gender bias in neuroticism measurement]. *Ciencias Sociales, 69*, 51-66.
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. En P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.
- Ellis, B., & Mead, A. (2000). Assessment of the measurement equivalence of a Spanish translation of the 16PF questionnaire. *Educational and Psychological Measurement, 60*, 787-807.
- Escorial, S. (2008). *Proceso de validación de un instrumento de evaluación psicológica para la medición de las dificultades del temperamento según el modelo de David Lykken*. [Validation of an assessment instrument for measuring temperament difficulties on the model of David Lykken.] (Tesis de doctorado sin publicar, Facultad de Psicología, Universidad Autónoma de Madrid, España).
- Escorial, S., & Navas, M. J. (2006). Análisis de la variable género en las escalas del EDTC mediante técnicas de funcionamiento diferencial de los ítems [Analysis of the Gender Variable in the EDTC Scales Using Differential Item Functioning Techniques]. *Psicothema, 18*(2), 319-325.
- Escorial, S., & Navas, M. J. (2007). Analysis of the Gender Variable in the Eysenck Personality

- Questionnaire Revised Scales Using Differential Item Functioning Techniques. *Educational and Psychological Measurement*, 67(6), 990-1001.
- Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement*, 61, 517-531.
- Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, 116 (3), 429-456.
- Ferrando, P. J., & Chico, E. (2001). The construct of sensation seeking as measured by Zuckerman's SSS-V and Arnett's AISS: a structural equation model. *Personality and Individual Differences*, 31, 1121-1133.
- Fidalgo, A. M. (1996). Funcionamiento diferencial de los ítems [Differential item functioning]. In J. Muñiz (Ed.), *Psicometría* (pp. 371-455). Madrid: Universitas.
- Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, 23, 267-269.
- Gelin, M., & Zumbo, B. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the Center for Epidemiologic Studies Depression Scale. *Educational and Psychological Measurement*, 63, 65-74.
- Gomá i Freixanet, M. (1995). Prosocial and antisocial aspects of personality. *Personality and Individual Differences*, 19, 125-134.
- Gray, J. M., & Wilson, M. A. (2007). A detailed analysis of the reliability and validity of the sensation seeking scale in a UK sample. *Personality and Individual Differences*, 42, 641-651.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.
- Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning*. Hillsdale: Lawrence Erlbaum Associates.
- Jensen, A. R. (1998). *The g factor*. Westport, Connecticut: Praeger.
- Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, 14, 329-349.
- Jorm, A. (1987). Sex differences in neuroticism: A quantitative synthesis of published research. *Australian and New Zealand Journal of Psychiatry*, 21, 501-506.
- Lange, R. Irwin, H. J. & Houran, J. (2000). Top-down purification of Tobacyk's Revised Paranormal Belief Scale. *Personality and Individual Differences* 29, 131-156.
- Leo, J. A., Van Dam, N. T., Hobkirk, A. L., & Earleywine, M. (2011). Examining bias in the impulsive sensation seeking (ImpSS) Scale using Differential Item Functioning (DIF) - An item response analysis. *Personality and Individual Differences* 50, 570-576.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, N. J.: Erlbaum.
- Lumsden, J. (1976). Test theory. En M. R. Rosenzweig & L.W. Porter (Eds.) *Annual Review of Psychology*. Palo Alto, CA: Annual Reviews Inc.
- Lykken, D. (1995). *The antisocial personalities*. New Jersey: Lawrence Erlbaum Associates, Inc.
- McDonald, R. P. & Mok, M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, 30, 23-40.
- Pérez, J., & Torrubia, R. (1986). Fiabilidad y validez de la versión española de la escala de búsqueda de sensaciones (Forma V) [Reliability and validity of the Spanish adaptation of the SSS-V]. *Revista Latinoamericana de Psicología*, 18, 7-22.
- Raju, N., van der Linden, W., & Fleer, P. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, 19, 353-368.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor test: results and

- implications. *Journal of Educational Statistics*, 4, 207-230.
- Reise, S., Smith, L., & Furr, M. (2001). Invariance on the NEO PI-R Neuroticism scale. *Multivariate Behavioral Research*, 36, 83-110.
- Roberti, J. W. (2004). A review of behavioral and biological correlates of sensation seeking. *Journal of Research in Personality*, 38, 256-279.
- Roth, M. & Hammelstein, P. (2003). *Sensation seeking - Konzeption, Diagnostik und Anwendung [Sensation Seeking-Conception, Measurement and Application]*, Hogrefe: Göttingen.
- Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel Hanzel Type I error performance. *Journal of Educational Measurement*, 32, 215-230.
- Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-240). Hillsdale, NJ: Lawrence Erlbaum.
- Smith, L. (2002). On the usefulness of item bias analysis to personality psychology. *Personality and Social Psychology Bulletin*, 28, 754-763.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item/test functioning (DIF/DTF) on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, 89, 497-508.
- Stout, W., & Roussos, L. (1999). *Dimensionality-based DIF/DBF package* [Computer software]. Urbana-Champaign: William Stout Institute for Measurement, University of Illinois.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361-370.
- Tanaka, J. S. & Huba, G. J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology*, 38, 197-201.
- Thissen, D. (1991). *MULTILOG: Multiple, categorical item analysis and tests scoring using item response theory*. Chicago: Scientific Software.
- Trimpop, R. M., Kerr, J. H., & Kirkcaldy, B. (1999). Comparing personality constructs of risk-taking behavior. *Personality and Individual Differences*, 26, 237-254.
- Waller, N. (1998). LINKDIF: Linking item parameters and calculating IRT measures of differential item functioning of items and tests. *Applied Psychological Measurement*, 22, 392.
- Zuckerman, M. (1979). *Sensation seeking: beyond the optimal level of arousal*. Hillsdale: Erlburn.
- Zuckerman, M. (1984). Sensation Seeking: A comparative approach to a human trait. *Behavioral and Brain Sciences*, 7, 413-471.
- Zuckerman, M. (1990). The psychophysiology of sensation seeking. *Journal of Personality*, 58, 313-345.
- Zuckerman, M. (1994). *Behavioral expressions and biosocial bases of sensation seeking*. New York: Cambridge University Press.
- Zuckerman, M. (2005). *Psychobiology of personality (Second edition, revised and updated)*. New York: Cambridge University Press.
- Zuckerman, M. (2007a). *Sensation seeking and risky behavior*. Washington, D.C.: American Psychological Association.
- Zuckerman, M. (2007b). The sensation seeking scale V (sss-v): Still reliable and valid. *Personality and Individual Differences*, 43, 1305-1307.
- Zuckerman, M., Eysenck, S.B.G., & Eysenck, H. J. (1978). Sensation seeking in England and America: Cross-cultural, age, and sex comparisons. *Journal of Consulting and Clinical Psychology*, 46, 139-149.
- Zuckerman, M., & Neeb, M. (1980). Demographic influences in sensation seeking and expressions of sensation seeking in religion, smoking and driving habits. *Personality and Individual Differences*, 1, 197-206.
- Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for*

studying DIF. Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia: Prince George, B. C.

- * Ítems que mostraron diferencias de sexo significativas según la información presentada en la tabla 3. En negrita los valores estadísticamente significativos.

Recibido: 21 de septiembre de 2015
Aprobado: 11 de mayo de 2016

