

# Adaptación y análisis psicométrico de la escala de deseabilidad social de Marlowe y Crowne

Pere J. Ferrando y Eliseo Chico  
Universidad Rovira i Virgili

El presente estudio describe la adaptación al castellano de la escala de deseabilidad social de Marlowe y Crowne (SDS) y estudia sus propiedades psicométricas desde el marco de la teoría de respuesta a los ítems (TRI). Se discute también la evolución del constructo de deseabilidad social y la relevancia y utilidad de la SDS. La versión adaptada se muestra como esencialmente unidimensional y posee propiedades psicométricas similares a las de la versión original. El modelo de TRI elegido se muestra apropiado y aporta información de interés respecto a las propiedades de la escala.

*A Spanish version of the Marlowe and Crowne's social desirability scale.* A Spanish adaptation of the Marlowe and Crowne's social desirability scale (SDS) based on the Item Response Theory (IRT) is presented. The historical evolution of both the social desirability construct and the SDS are revised. The Spanish version of the SDS behaves as essentially unidimensional and has psychometric properties similar to those of the original version. The IRT model used to scale the SDS fits well the data and provides useful information about the properties of both the items and the total scale.

Es bien conocido que los tests de personalidad o de rendimiento típico muestran, en general, peores propiedades psicométricas que los tests de capacidad o de rendimiento máximo (véase e.g. Anastasi, 1973 o Cronbach, 1985). Una posible explicación a este fenómeno es la de que existen una serie de determinantes, aparte del contenido, que pueden influir en la respuesta a un ítem de personalidad (Jackson y Messik, 1958, Fiske y Butler, 1963; Jackson, 1973; Edwards, 1990; Viswesvaran y Ones, 1999). De entre estos potenciales determinantes, el que ha recibido más atención ha sido la deseabilidad social (DS).

El interés acerca de la DS se inicia en los años 30 con los trabajos de Bernreuter (1933), alcanza su máximo apogeo entre los años 1955 y 1965 y parece que vuelve a reavivarse desde principios de los 90. La cantidad de literatura generada por el tópico es enorme y su revisión queda fuera de los objetivos de este trabajo. Nos limitaremos pues a dar una visión de algunos de los aspectos más polémicos así como algunas referencias para el lector interesado.

El punto de partida del constructo DS es el supuesto de que algunas de las alternativas de respuesta en ciertos reactivos de personalidad son socialmente más convenientes o deseables que otras, por lo que algunos sujetos pueden tender a elegir las independientemente de cual sea su nivel en el rasgo que el reactivo pretende medir (Edwards, 1953, 1957, 1990). Más en detalle, la teoría inicial planteaba que ciertos sujetos podían dar, voluntariamente, una imagen distorsionada de sí mismos respondiendo siempre a

las alternativas socialmente más deseables de los reactivos. Lógicamente, se planteaba también que esta tendencia a 'disimular' o a 'quedar bien' se acentuaría cuando la motivación para hacerlo fuese alta, por ejemplo en situación de selección para un puesto de trabajo (Michaelis y Eysenck, 1971). Como consecuencia de este planteamiento, el esfuerzo de algunos constructores de tests se encaminó a desarrollar escalas de DS o de 'mentira' que permitiesen detectar a los sujetos que distorsionaban sus respuestas.

Si bien la teoría inicial era plausible, algunos autores se plantearon hasta qué punto la distorsión por DS era un problema real en personalidad. Esto dio lugar a una polémica muy polarizada en la que algunos autores (e.g. Edwards, 1957, Jackson, 1973) consideraban que los sujetos respondían más a la deseabilidad de un cuestionario que a su contenido, en tanto que otros autores (e.g. Block, 1965; Rorer, 1965) consideraban que el papel de la DS era muy pequeño o incluso inexistente. Cabe decir que con el tiempo las posturas se fueron suavizando un tanto (Block, 1990).

La investigación empírica con las escalas de DS construidas en los años 50 y 60 fue mostrando ciertos resultados de interés. En muchos casos estas escalas se administraban con fines de investigación básica, es decir bajo condiciones que no suscitaban ninguna motivación para el fingimiento, pero aún y así: (a) las puntuaciones medias se situaban en el punto medio de la escala, (es decir no había efecto suelo), (b) las puntuaciones tenían bastante varianza y (c) la fiabilidad era bastante alta. Todo esto sugería que, al menos en estas condiciones, estas escalas estaban midiendo un rasgo estable (véase Eysenck y Eysenck, 1976 para una revisión del tema). Debido a estos resultados, la DS dejó de verse tan solo como una tendencia a dar respuestas voluntariamente distorsionadas para pasar a verse como un rasgo relativamente estable y consistente de personalidad. Dicho rasgo ha recibido diferentes nombres, Crowne y Marlowe (1964), por ejemplo, lo denominan 'Necesidad de Aprobación Social', en tanto que Furnham (1986) lo

denomina 'Conformismo Social', pero el contenido que subyace a las diferentes denominaciones es esencialmente el mismo.

Con este cambio de enfoque, por supuesto, también cambia la polémica. Ya no se trata tanto de estudiar hasta qué punto las características de los ítems invitan a la disimulación, como de estudiar hasta qué punto un rasgo de personalidad está relacionado (o contaminado, según se mire) con el rasgo de DS. En términos psicométricos el problema no es ya un problema de sesgos de respuesta sino de validez discriminante (Hogan y Nicholson, 1988). Como es de suponer, el aspecto que actualmente suscita más controversia es la forma en que deben interpretarse las correlaciones entre las medidas de diferentes rasgos de personalidad y las medidas de DS (Hogan y Nicholson, 1988). Lo que está fuera de toda duda, en cambio, es que tales correlaciones existen y que son bastante substanciales en algunos casos (Borkenau y Ostendorf, 1992).

Quedan aún muchos más aspectos a resolver en el tema DS. Por ejemplo, como era de esperar, algunos autores sugieren que la DS no es un rasgo unitario sino multidimensional (Paulhus, 1984). Tampoco está claro cómo se relacionan los enfoques antiguo y moderno. Eysenck y Eysenck (1976) sugieren que las escalas de DS miden un rasgo estable cuando se administran en condiciones neutrales o de baja motivación, pasando a medir disimulación en situaciones de presión. Sin embargo, las cosas no son tan simples y parece ser que la forma en que los niveles en el rasgo y la disimulación influyen sobre las puntuaciones en DS bajo distintas condiciones de motivación son harto complejas (Furham, 1986; Cowles, Darling y Skanes, 1992; Elliott, Lawty-Jones y Jackson 1996). Finalmente tampoco está clara cual es la validez predictiva de las puntuaciones en DS considerada como rasgo de personalidad (Ones, Viwesvaran y Reiss, 1996).

Para terminar esta revisión, necesariamente incompleta, los autores recomendamos al lector interesado la lectura del ejemplar de febrero de 1990 de la revista 'American Psychologist', donde se publicaron una serie de debates acerca de la DS escritos, en su mayor parte, por los autores más relevantes en la historia del constructo.

#### La Escala de DS de Marlowe y Crowne

La escala de deseabilidad social de Marlowe y Crowne (conocida en inglés con las siglas SDS) fue desarrollada por estos autores para superar las limitaciones encontradas en la, aún popular hoy día, escala de DS de Edwards (1957). Los 39 ítems de la escala de Edwards procedían del MMPI y, por tanto, muchos de ellos tenían contenido psicopatológico, por lo que la principal crítica de Crowne y Marlowe era que la escala de Edwards confunde DS con psicopatología. Crowne y Marlowe (1960) decidieron elaborar una escala totalmente nueva con reactivos que incluyesen un mínimo de contenido patológico. Los 33 ítems de la SDS difieren principalmente de los de Edwards en dos aspectos: (a) no hacen referencia a aspectos psicopatológicos y (b) representan conductas culturalmente sancionadas o aprobadas pero con poca probabilidad de ocurrencia. Para una comparación metodológicamente rigurosa entre la escala de Edwards y la SDS, el lector puede consultar O'Grady (1988).

La SDS tuvo una inmediata aceptación y aún hoy día, a sus casi 40 años, sigue siendo el instrumento más popular y utilizado para medir la deseabilidad social (Reynolds, 1992; Schmitt y Steyer, 1993). La cantidad de literatura referida a la escala es enorme e im-

posible de resumir en un artículo como éste. Nos limitaremos, por tanto a revisar algunos estudios que hacen referencia a sus propiedades psicométricas así como a adaptaciones a distintas culturas.

Las puntuaciones de la SDS tienen propiedades psicométricas 'básicas' que pueden considerarse aceptables dado el tipo de instrumento de que se trata. Las fiabilidades estimadas mediante índices de consistencia interna se mueven entre .75 y .85 (Strahan y Gerbasi, 1972; Reynolds, 1982; Ballard, 1992; Borkenau y Ostendorf, 1992) y la estabilidad temporal tras un intervalo de un mes se estimó en .89 (Crowne y Marlowe, 1960). Respecto a sus debilidades, las principales críticas son de dos tipos: (a) los bajos índices de discriminación de bastantes de sus ítems y (b) falta de unidimensionalidad.

La primera crítica puede deducirse del hecho de que la escala sólo llega a fiabilidades mínimamente aceptables aún siendo bastante larga (33 ítems). Los estudios en los que se fundamenta esta primera crítica son todos del mismo tipo. Se lleva a cabo un análisis en componentes principales y se estudian las cargas de los ítems en el primer componente no rotado. Los resultados pueden resumirse como sigue. Stratham y Gerbasi (1972) encuentran un valor medio de carga de 0.35 y Ballard (1992) de 0.33. Ballard, Crino y Rubinfeld (1988) encuentran que sólo 16 de los 33 reactivos tienen cargas superiores a 0.40, en tanto que Reynolds (1982) reporta que sólo 11 reactivos cumplen esta condición. La solución propuesta por todos estos autores es la misma: eliminar los reactivos con cargas bajas y llegar con esto a una versión reducida de la escala, más consistente y que apenas pierde fiabilidad con respecto a la escala original. La crítica que puede hacerse a todos ellos es la misma: ninguno lleva a cabo un estudio de validación cruzada.

La segunda crítica se fundamenta en dos tipos de estudios distintos. Los estudios del primer tipo son análisis factoriales a nivel de ítem (Millham, 1974; Ramanaiah y Martin, 1980), y sugieren una estructura de la SDS en dos factores: un factor de 'atribución' (tendencia a adjudicarse conductas socialmente deseables) y un factor de 'negación' (tendencia a negar conductas indeseables). Es de notar que estos factores están definidos por los ítems redactados en sentido positivo y negativo respectivamente en la escala. Respecto al segundo tipo de estudios, se basan en análisis factoriales a nivel de escala (Frenken-Brunswick, 1939; Meelk y Hataway, 1946; Borkenau y Ostendorf, 1992, Fischer y Frick, 1993) y sugieren una estructura bidimensional del constructo DS, con un factor de 'autoengaño' (la persona cree que su respuesta favorable es correcta) y un factor de 'manejo de la imagen' (la persona manipula conscientemente la imagen que desea mostrar). La escala de Edwards sería esencialmente una medida del primer factor, en tanto que las escalas de mentira de los cuestionarios de Eysenck serían medidas bastante puras del segundo factor. La SDS, por último sería una mezcla que mediría ambos factores (Paulhus, 1984; Borkenau y Ostendorf, 1992).

La SDS ha sido traducida y adaptada a diversos idiomas (e.g. Schmitt y Steyer, 1993). Con vistas al presente trabajo, es de interés citar que existe una versión de la SDS en español desarrollada en México por Castro, Moya y Orozco (1986) y estudiada por Lara-Cantú (1988, 1990). Los autores revisamos esta versión y decidimos que no podía utilizarse directamente en nuestro país porque contenía términos y expresiones que aquí resultan muy poco habituales. Por esta razón decidimos llevar a cabo la adaptación directamente del original. No obstante los resultados obtenidos por Lara-Cantú se compararán con los resultados que se obtengan en el presente estudio.

## Objetivos

Los objetivos del presente trabajo son de dos tipos, substantivos y metodológicos. Desde el punto de vista substantivo, los autores pretendemos adaptar al castellano uno de los instrumentos más utilizados en medición de la personalidad y estudiar sus propiedades psicométricas. Juzgado como un estudio sustantivo, el interés del presente trabajo dependerá de la importancia y utilidad de la escala que se adapta. La importancia de la SDS parece estar fuera de toda duda. Su utilidad (con las debidas reservas) puede decirse que es doble: (a) administrada en condiciones de alta motivación, puede servir para detectar a los sujetos que tienden a distorsionar sus respuestas a fin de presentarse a sí mismos en la forma más favorable posible, y (b) administrada en condiciones neutrales pretende medir un rasgo de conformismo social y permite evaluar la validez discriminante de otras escalas de personalidad con referencia a este rasgo.

Desde el punto de vista metodológico el presente trabajo pretende mostrar que es posible utilizar la Teoría de Respuesta a los Ítems (TRI) para analizar y escalar un test convencional de personalidad. En una reciente revisión, Steinberg y Thissen (1995) consideraban que la TRI se había utilizado muy poco en personalidad, porque dicha teoría es especialmente adecuada para el escrutinio detallado de pequeños conjuntos de ítems, en tanto que la práctica habitual en personalidad es la de utilizar tests largos (de más de 20 ítems) y, en general, poco consistentes. En este trabajo se quiere demostrar que la TRI es también útil para analizar y escalar un test de personalidad con estas características y que proporciona información de interés acerca de las propiedades de los ítems.

## Método

### *Participantes*

La muestra analizada en este trabajo está formada por 847 estudiantes universitarios (592 mujeres, 255 hombres, edad mediana de 20 años) de las facultades de Psicología y Pedagogía, y de las escuelas universitarias de Magisterio y Trabajo Social de la universidad 'Rovira i Virgili' (Tarragona). Todos ellos participaron voluntariamente y no recibieron ningún tipo de compensación por ello. Las condiciones bajo las que participaron pueden considerarse pues como totalmente neutras en términos de motivación.

### *Instrumentos*

La escala original de Marlowe y Crowne (1960, Tabla 1) fue traducida y adaptada por uno de los autores del presente trabajo con la colaboración del profesor Gabriel Molina, de la universidad de Valencia, y posteriormente fue revisada independientemente por dos profesores de inglés nativos que llevaban más de 15 años viviendo en nuestro país. La prueba piloto fue administrada a un reducido grupo de estudiantes y algunos ítems se modificaron ligeramente de acuerdo con las sugerencias de los participantes. Por ejemplo, el ítem 27: 'Nunca emprendo un viaje largo sin revisar el coche' fue cambiado a: 'Nunca emprendo un viaje largo sin revisar el coche (moto, bici, etc.)' ya que muchos de los estudiantes no tenían coche. La versión resultante tras estas ligeras modificaciones es la que se empleó en el estudio que se presenta. El protocolo del test se presenta en el apéndice 1.

## *Procedimiento*

La administración de la escala se llevó a cabo dentro del aula y en el horario académico de los estudiantes, siempre con el margen de tiempo necesario para permitir a los participantes contestar sin premura de tiempo. La SDS fue administrada por uno de los autores, quien destacó en todo momento la voluntariedad en la realización de la prueba, la finalidad del trabajo (investigación básica) y el anonimato de los resultados individuales.

## Análisis y resultados

### *Análisis Preliminares*

En la primera etapa se estimó la fiabilidad de las puntuaciones en la SDS y se estudió su distribución de frecuencias. La distribución de las puntuaciones era acampanada y bastante simétrica con media  $M=15.83$  (son 33 ítems) y desviación típica  $SD=5.15$ , valores bastante similares a los obtenidos con la versión original de la escala. Crowne y Marlowe (1960), por ejemplo reportan  $M=13.72$  y  $SD=5.78$  y Reynolds (1992)  $M=15.00$  y  $SD=5.91$ . En los estudios realizados en México, en cambio, se obtienen medias bastante más altas. Lara-Cantú, reporta una media de 19.76 y  $SD=5.17$  en el estudio de 1988 y  $M=19.10$  y  $SD=5.57$  en el estudio de 1990. Estos resultados apoyan la evidencia empírica más general de que los latinoamericanos tienden a puntuar bastante más alto que los europeos o norteamericanos en DS (Shultz y Chávez, 1994). Por supuesto que es arriesgado hacer comparaciones transculturales, pero dado que todos los estudios descritos en este apartado se basaron en muestras de estudiantes universitarios, estos resultados sugieren que los estudiantes españoles que participaron en el presente estudio se parecen más a sus homólogos europeos o norteamericanos que a sus homólogos mexicanos en lo que refiere a sus niveles de DS.

El valor estimado para el cociente de fiabilidad (alpha) fue de 0.78, el mismo que el obtenido por Lara-Cantú (1990) utilizando la adaptación mexicana, y similar a las estimaciones a las que se llega cuando se utiliza la escala original, donde se obtienen valores tales como .73 (Strahan y Gerbasi, 1972), .82 (Reynolds, 1982) o .75 (Ballard, 1992).

### *Evaluación de la Dimensionalidad de la SDS*

La evaluación previa de la dimensionalidad de la SDS es importante por dos razones. En primer lugar para evaluar el fundamento de las críticas que sugieren una estructura bifactorial. En segundo lugar, porque el modelo de TRI que se va a utilizar en la siguiente etapa asume unidimensionalidad y, caso de no cumplirse este supuesto, se puede llegar a estimaciones distorsionadas de los parámetros de los ítems (véase Cuesta y Muñiz, 1999).

McDonald y Ahlawat (1974) demostraron que el uso del análisis factorial (AF) lineal en el caso de reactivos binarios podía dar lugar a factores artificiales si las regresiones ítem-factor eran no lineales. La TRI se basa, precisamente, en el supuesto de que la regresión ítem-factor en el caso binario es no lineal, por lo que cabe esperar que el uso del AF lineal sobre conjuntos de ítems binarios lleve a evidencia espuria de multidimensionalidad, es decir, que para conseguir un buen ajuste sea necesario extraer más factores de los que son propiamente factores de contenido.

Para minimizar este problema en el presente caso se utilizó el modelo de AF no lineal de McDonald (1967). El principio básico de este modelo es el mismo que el del AF lineal: el de que las covarianzas residuales se anulan tras extraer el número apropiado de factores. Sin embargo, este modelo ajusta las regresiones ítem factor mediante un polinomio cúbico, por lo que es de esperar que no aparezcan factores espurios debidos a la no linealidad en la evaluación dimensional. El programa utilizado para ajustar el modelo fue NOHARM (Fraser y McDonald, 1988), un programa que utiliza análisis armónico y que se basa en el criterio de mínimos cuadrados.

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), los autores hemos utilizado también la versión para mínimos cuadrados del índice normado  $g$  de bondad de ajuste propuesto por Tanaka y Huba (1985). Basándonos en la evidencia empírica discutida antes, decidimos evaluar modelos de uno y dos factores.

Los resultados obtenidos fueron los siguientes. Para el modelo de un factor común, la RMSR fue de 0.008 y  $g$  de 0.880. Para el modelo de dos factores, estos valores fueron de 0.007 y de 0.912 respectivamente. Así pues: (a) el modelo de un sólo factor muestra un ajuste razonablemente bueno y (b) la mejora en la bondad de ajuste que se produce al pasar de uno a dos factores es bastante pequeña.

Hay otros indicios que apuntan a que la SDS se muestra como esencialmente unidimensional en este estudio. Así, cuando se examinó la solución rotada en dos factores (Promax) se vio que el segundo factor estaba determinado por dos ítems: el ítem 5 'Algunas veces dudo de mi habilidad para triunfar en la vida', y el ítem 10 'En algunas ocasiones he renunciado a hacer algo porque pensaba que me faltaba habilidad'. La similitud en el contenido de estos dos ítems sugiere que este débil segundo factor es un factor espurio que refleja un problema de residuales correlacionados (para una explicación substantiva de los residuales correlacionados en personalidad véase Fiske, 1978). En una eventual versión depurada de la escala sería de interés considerar la posible eliminación de uno de estos dos ítems. En todo caso, sin embargo, parece claro que los presentes datos no apoyan una estructura en dos factores interpretables.

#### *Propuesta de un Modelo de TRI y Evaluación de su Ajuste*

La revisión de la literatura muestra que el modelo TRI más utilizado para reactivos binarios de personalidad es el modelo logístico de dos parámetros (ML2P, Reise y Waller, 1990; Waller, Tellegen, McDonald y Likken, 1996; Finch y West, 1997; Reise, 1999). Esto es así por dos razones: (a) los ítems de personalidad suelen variar bastante en cuanto a sus niveles de discriminación (lo que excluye en muchos casos el uso del modelo de Rasch) y (b) el parámetro de adivinación habitualmente carece de importancia en este contexto, lo que hace innecesario el uso del modelo de tres parámetros. Por estas razones, los autores decidimos calibrar los ítems mediante el ML2P. Cabe notar, sin embargo, que los resultados de la calibración sólo deben interpretarse si el modelo proporciona un buen ajuste a los datos.

El ML2P fue ajustado mediante el programa BILOG-3 (Mislevy y Bock, 1990). El procedimiento de estimación fue el de máxima verosimilitud marginal a posteriori y la distribución latente a priori se especificó como normal tipificada.

La evaluación del ajuste en el caso de un test largo como en el presente caso debe hacerse ítem a ítem. Para cada ítem las frecuencias observadas en distintos niveles del rasgo se comparan con las frecuencias esperadas de acuerdo al ML2P por medio de un estadístico basado en la distribución ji-cuadrado (Mislevy y Bock, 1990). Para cada uno de los 33 ítems, los valores del estadístico, los correspondientes grados de libertad y la probabilidad asociada se presentan en las tres últimas columnas de la tabla 1.

Como muestra la tabla 1, el ajuste del modelo es bueno en general y, al nivel convencional del 5%, habría que concluir que sólo tres ítems muestran un ajuste deficiente: el ítem 25, el 5 y el 18. Los autores no hemos encontrado una explicación para el mal ajuste de los dos primeros. Respecto al 18, el mal ajuste, unido al hecho de que es el ítem con un valor más bajo del índice de discriminación sugiere problemas de ambigüedad. En primer lugar el enunciado es enrevesado y quizás sería preferible redactarlo en sentido positivo, por ejemplo: 'Me resulta bastante fácil relacionarme con gente escandalosa y detestable'. Además, no está clara la dirección de la DS en la respuesta. Por ejemplo contestar 'verdadero' puede entenderse tanto como socialmente deseable (tolerante, capaz de relacionarse con todo el mundo) como indeseable (hacerse sospechoso de frecuentar malas compañías). Cabe destacar que el mal funcionamiento de este ítem no es una característica exclusiva de nuestra adaptación, ya que en los estudios factoriales basados en la escala original, este ítem tiene generalmente cargas prácticamente nulas en el primer factor (Reynolds, 1992).

*Tabla 1*  
Calibración de la SDS mediante el modelo logístico de 2 parámetros

ITEM	a	b	$\chi^2$	g.l.	P
15	.908	-.215	10.5	7	.161
16	.823	-.731	3.2	7	.864
6	.778	2.026	6.0	6	.427
3	.769	2.043	3.8	7	.804
12	.766	.589	10.2	9	.333
33	.763	.188	4.5	8	.807
19	.747	-.376	7.4	8	.499
13	.744	-.192	4.0	8	.857
26	.680	-.174	8.3	8	.401
30	.631	-1.154	7.6	8	.470
20	.609	-2.446	5.2	6	.518
29	.594	1.068	3.6	9	.932
2	.587	-1.617	10.1	7	.181
21	.581	-.054	7.6	9	.571
9	.579	.747	3.3	9	.950
24	.572	-1.548	10.1	7	.180
28	.569	.513	2.9	9	.965
11	.568	1.837	3.0	8	.936
22	.561	2.482	10.6	7	.157
4	.559	1.585	8.9	8	.353
8	.557	.877	11.5	9	.243
25	.464	-2.428	15.1	7	.034
17	.447	-2.257	3.8	8	.876
14	.444	2.143	7.8	9	.559
31	.432	.791	14.7	9	.098
23	.379	1.168	3.8	9	.924
32	.377	-2.304	10.8	8	.210
5	.339	1.714	18.8	9	.026
27	.339	-.811	7.4	9	.600
10	.301	1.473	5.0	9	.835
1	.273	-.046	6.3	9	.708
7	.181	.181	11.9	9	.216
18	.129	2.294	17.2	9	.045

### Calibración de los Ítems

Las tres primeras columnas de la tabla 1 muestran: la posición del ítem en la escala (ver apéndice 1), el índice de discriminación  $a$  y el índice de dificultad  $b$ . Para facilitar la interpretación, los ítems han sido ordenados de mayor a menor según el índice de discriminación.

Comparando los resultados de la tabla con el enunciado de los ítems en el apéndice, puede verse que los reactivos con valores más altos de discriminación tienen enunciados muy claros y refieren a situaciones muy generales, aplicables prácticamente a cualquier muestra. Por contra los ítems con valores más bajos de discriminación, o bien son poco claros (ítem 10 o ítem 18) o son poco aplicables a esta muestra (ítem 1 e ítem 7). En cuanto a los índices de dificultad, son bastante directos de interpretar. Por ejemplo, el ítem más 'difícil' es el 22 y, en efecto, hace falta un nivel muy alto de DS para contestar 'falso' al enunciado: 'A veces insisto en hacer las cosas a mi manera'.

Reise y Waller (1990) consideran que, en el caso de cuestionarios de personalidad, los valores de  $b$  suelen moverse entre  $-2$  y  $+2$ , y los de  $a$  entre  $0.50$  y  $1.50$ . En el presente estudio, el rango de valores de  $b$  es de  $-2.446$ ;  $2.482$  (media  $0.223$ ) y el de  $a$  es de  $0.129$ ;  $0.908$  (media  $0.547$ ). El rango de valores de  $b$  es pues algo más amplio de lo que es habitual en este tipo de tests, lo cual es ventajoso para un test destinado a la población general ya que indica que se podrá medir con razonable precisión en un amplio intervalo de valores del rasgo. Respecto a los índices de discriminación, sus valores son más bajos de lo habitual y esto no es una característica deseable, ya que indica que los reactivos no miden con demasiada precisión. Sin embargo, este resultado no invalida a la escala, tan sólo indica que ésta deberá ser larga si quiere llegarse a una precisión aceptable en la medida. En la próxima sección se discute este tema más en detalle.

### Función de Información

Uno de los autores escribió un sencillo programa en MATLAB para representar la función de información del test (FIT) correspondiente a diferentes niveles del rasgo a medir (véase Muñiz, 1990, p.97 para la determinación de la FIT). El gráfico correspondiente a un rango de valores entre  $-3$  y  $+3$  se presenta en la figura 1.

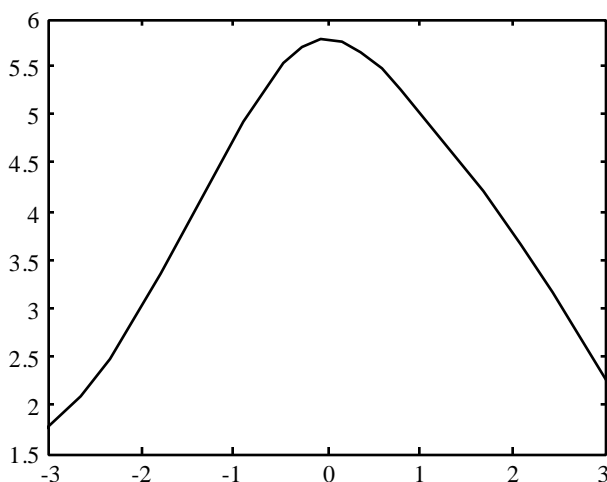


Figura 1. Función de información del test

La figura 1 muestra varias características de interés. En primer lugar la escala rinde su máxima información en torno a un nivel del rasgo de cero (es decir en torno a la media si se asume una distribución latente normal estandarizada). Esto indica que el nivel del test es adecuado para el grupo considerado en este estudio, lo cual tiende a reforzar la hipótesis de Eysenck y Eysenck (1976) comentada anteriormente. En segundo lugar la FIT describe una curva relativamente simétrica y poco apuntada, lo cual indica que el test puede funcionar bien para un rango de valores del rasgo bastante amplio, tanto por encima como por debajo de la media. Este resultado, por supuesto, está directamente relacionado con el amplio rango de valores de los índices de dificultad al que antes nos hemos referido.

Como aspectos negativos, cabe notar que los valores generales de la FIT son bastante bajos. El máximo del valor de la función es de  $5.8$  y se obtiene en torno a un nivel de cero. Usando la bien conocida relación entre la FIT y el error típico de medida (véase e.g. Muñiz, 1990, p. 93), se deduce que incluso en el mejor de los casos el error típico es tan grande como  $0.41$ , lo cual indica claramente que el test no tiene una gran precisión en la estima del rasgo sea cual fuere el nivel de éste. Este resultado podría parecer sorprendente ya que el test es relativamente largo (33 ítems). Sin embargo, en el modelo de dos parámetros, la FIT depende fundamentalmente de dos factores: la longitud del test y la magnitud de los índices de discriminación, por lo que el resultado obtenido no hace más que reflejar un problema endémico en los tests de personalidad: el de los bajos valores de los índices de discriminación (Steinberg y Thissen, 1995). Este problema es más acusado en el presente caso donde, como hemos visto, los índices de discriminación son bajos incluso para los estándares de los tests de personalidad.

### Discusión y Conclusiones

Tal como se pretendía en los objetivos, el presente trabajo demuestra que es posible utilizar la TRI en el análisis de un cuestionario de personalidad relativamente largo y obtener de este análisis información interesante tanto en lo que refiere al comportamiento de los ítems individuales como al del test total. Esta conclusión ha sido también obtenida en trabajos anteriores que utilizan instrumentos típicos en medición de la personalidad (Ferrando, Varea y Lorenzo, 1999).

Con algunas limitaciones, la presente adaptación de la SDS posee propiedades psicométricas aceptables. Sus puntuaciones miden esencialmente una sola dimensión y la mayoría de sus ítems son escalables de acuerdo al modelo logístico de dos parámetros. Su nivel es adecuado para el grupo normativo utilizado y, de acuerdo con su curva de información, parece ser utilizable en un amplio rango de niveles en el rasgo a medir.

En cuanto a sus limitaciones, éstas se refieren principalmente a los bajos índices de discriminación de bastantes de sus ítems, una limitación que, como hemos visto, no cabe atribuir a la adaptación sino que es propia ya de la escala original. Sin embargo, en base al presente análisis, no parece recomendable construir formas reducidas (y más homogéneas) de la escala ya que los índices  $a$  son bajos en general y, si la escala se redujese notablemente, perdería demasiada precisión en la medida. Quizás lo ideal sería 'pulir' la escala eliminando los pocos ítems que tienen valores de discriminación muy bajos (digamos por debajo de  $0.30$ ).

No hay duda de que el presente trabajo deja pendientes una serie de puntos que pueden ser objeto de futura investigación. Desde un punto de vista psicométrico sería de interés estudiar la validez predictiva de la escala respecto a criterios externos, por ejemplo en situaciones de selección y, asimismo, sería interesante elaborar baremos en nuestro país y llevar a cabo investigaciones de tipo transcultural. Desde un punto de vista más experimental, sería de interés estudiar cómo se comportan las puntuaciones de la escala bajo diferentes condiciones de presión o motivación.

Para algunos autores (e.g. Kline, 1986, Nunnally, 1987), el proceso de validación de cualquier test de personalidad debe incluir siempre una prueba empírica que indique que el test no se ve afectado por la variable de deseabilidad social, siendo la prueba que habitualmente se sugiere la de una correlación no significativa entre las puntuaciones del test bajo estudio y las puntuaciones en la SDS (Kline, 1986). El trabajo que aquí se presenta permite realizar esta prueba de validación discriminante en el caso de tests desarrollados o adaptados en castellano. Su utilidad, en última instancia, dependerá principalmente del futuro uso que se haga de la presente adaptación.

### Apéndice 1. Versión española de la SDS

#### ESCALA D.S. DE MARLOWE Y CROWNE ADAPTACIÓN ESPAÑOLA

Apellidos y nombre ..... Edad: ..... Sexo: .....

A continuación verás una serie de frases que están relacionadas con actitudes personales. Lee atentamente cada una de ellas y decide si tu forma habitual de ser se parece (V) o no (F) al contenido de la frase. No dejes ninguna frase sin responder.

1	Antes de votar me informo detalladamente de la capacidad de todos los candidatos	V	F
2	Nunca dudo en dejar lo que estoy haciendo para ayudar a alguien con problemas	V	F
* 3	A veces me cuesta ponerme a trabajar si no me encuentro con ánimos	V	F
4	Nunca me ha caído nadie realmente mal	V	F
* 5	Algunas veces dudo de mi habilidad para triunfar en la vida	V	F
* 6	A veces estoy descontento cuando no puedo hacer las cosas a mi manera	V	F
7	Siempre soy muy cuidadoso con mi manera de vestir	V	F
8	En casa, me comporto tan bien en la mesa como cuando voy a un restaurante	V	F
* 9	Si pudiera entrar en una sala de cine sin pagar y estuviera seguro de que no me vieran, probablemente lo haría	V	F
* 10	En algunas ocasiones he renunciado a hacer algo porque pensaba que me faltaba habilidad	V	F
* 11	A veces me gusta chismorrear un poco	V	F
* 12	Ha habido veces en que he tenido sentimientos de rebeldía contra personas con autoridad aún sabiendo que ellos tenían la razón	V	F
13	Independientemente de con quién esté hablando, siempre escucho atentamente	V	F
* 14	Alguna vez me «he hecho el loco» para quitarme a alguien de encima	V	F
* 15	En alguna ocasión me he aprovechado de alguien	V	F
16	Cuando cometo un error siempre estoy dispuesto a admitirlo	V	F
17	Siempre intento practicar lo que predico	V	F
18	No encuentro particularmente difícil relacionarme con gente escandalosa y detestable	V	F
* 19	A veces trato de vengarme en lugar de perdonar y olvidar lo que me han hecho	V	F
20	Cuando no sé algo no me importa admitirlo	V	F
21	Siempre soy cortés, aun con gente desagradable	V	F
* 22	A veces insisto en hacer las cosas a mi manera	V	F
* 23	En algunas ocasiones siento que soy un manazas	V	F
24	Nunca he dejado que alguien fuera castigado por cosas que había hecho yo	V	F
25	Nunca me enfado cuando me piden que devuelva algún favor que me han hecho	V	F
26	Nunca me irrito cuando la gente expresa ideas muy distintas de las mías	V	F
27	Nunca emprendo un viaje largo sin revisar el coche (moto, bici, etc.)	V	F
* 28	En algunas ocasiones me he sentido bastante celoso de la buena fortuna de los demás	V	F
29	Aún no he tenido nunca la necesidad de decirle a alguien que me dejara en paz	V	F
* 30	A veces me irrita la gente que me pide favores	V	F
31	Nunca me ha parecido que me castigaran sin motivo	V	F
* 32	A veces pienso que cuando la gente tiene mala suerte es porque se lo merece	V	F
33	Nunca he dicho deliberadamente nada que pudiera herir los sentimientos de alguien	V	F

#### MUCHAS GRACIAS POR SU COLABORACIÓN

\* Los ítems marcados con asterisco se puntúan en forma revertida.

## Referencias

- Anastasi, A. (1973). *Tests psicológicos*. Madrid: Aguilar.
- Ballard, R. (1992). Short forms of the Marlowe-Crowne social desirability scale. *Psychological Reports*, 71, 1.155-1.160.
- Ballard, R.; Crino, M.D. y Rubenfeld, S. (1988). Social desirability response bias and the Marlowe-Crowne social desirability scale. *Psychological Reports*, 63, 227-237.
- Bernreuter, R.G. (1933). Validity of the personality inventory. *Personality Journal*, 11, 383-386.
- Block, J. (1965). *The challenge of response sets*. New York: Appleton.
- Block, J. (1990). More remarks on social desirability. *American Psychologist*, 45, 9.
- Borkenau, P. y Ostendorf, F. (1992). Social desirability scales as moderator and suppressor variables. *European Journal of Personality*, 6, 199-214.
- Castro, M.E.; Maya, M.A. y Orozco, C. (1986). Normas y estructura factorial de las respuestas de la población estudiantil de la república mexicana a dos escalas: escala de responsabilidad intelectual y académica y escala de necesidad de aprobación social. *Salud Mental*, 9, 65-71.
- Cowles, M.; Darling, M. y Skanes, A. (1992). Some characteristics of the simulated self. *Personality and Individual Differences*, 13, 501-510.
- Cronbach, L.J. (1985). *Fundamentos de la exploración psicológica*. Madrid: Biblioteca Nueva.
- Crowne, D.P. y Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24, 349-354.
- Crowne, D.P. y Marlowe, D. (1964). *The approval motive: studies in evaluative dependence*. New York: Wiley.
- Cuesta, M. y Muñiz, J. (1999). Robustness of item response logistic models to violations of the unidimensionality assumption. *Psicothema*, 11, 175-182.
- Edwards, A.L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology*, 37, 90-93.
- Edwards, A.L. (1957). *The social desirability variable in personality assessment and research*. New York: Dryden.
- Edwards, A.L. (1990). Construct validity and social desirability. *American Psychologist*, 45, 287-289.
- Elliot, S, Lawty-Jones, M. y Jackson, C. (1996) Effects of dissimulation on self-report and objective measures of personality. *Personality and Individual Differences*, 21, 335-343.
- Eysenck, H.J. y Eysenck, S.B.G. (1976). *Psychoticism as a dimension of personality*. New York: Crane, Russak & Company.
- Ferrando, P.J., Varea, M.D. y Lorenzo, U. (1999). Evaluación psicométrica del cuestionario de ansiedad y rendimiento (CAR) en una muestra de escolares. *Psicothema*, 11, 225-236.
- Finch, J.F. y West, S.G. (1997). The investigation of personality structure: statistical models. *Journal of Research in Personality*, 31, 439-485.
- Fischer, D.G. y Fick, C. (1993). Measuring social desirability: short forms of the Marlowe-Crowne social desirability scale. *Educational and Psychological Measurement*, 53, 417-424.
- Fiske, D.W. (1978). *Strategies for personality research*. San Francisco: Jossey-Bass
- Fiske, D.W. y Butler, J.M. (1963). The experimental conditions for measuring individual differences. *Educational and Psychological Measurement*, 23, 249-266.
- Fraser, C. y McDonald, R.P. (1988). NOHARM: least squares item factor analysis. *Multivariate Behavioral Research*, 23, 267-269.
- Frenkel-Brunswick, E. (1939). Mechanisms of self-deception. *Journal of Social Psychology*, 10, 409-420.
- Furnham, A. (1986) Response bias, social desirability and dissimulation. *Personality and Individual Differences*, 7, 385-400.
- Hogan, R. y Nicholson, R.A. (1988). The meaning of personality test scores. *American Psychologist*, 43, 621-626.
- Jackson, D.N. (1973). Structured personality assessment. En B.B. Wolman (ed.) *Handbook of general Psychology* (pp. 775-792). Englewood Cliffs: Prentice Hall.
- Jackson, D.N. y Messick, S.J. (1958). Content and style in personality assessment. *Psychological Bulletin*, 55, 243-252.
- Kline, P. (1986). *A handbook of test construction*. Methuen: London.
- Lara-Cantú, M.A. (1990). Validez y confiabilidad de la escala de deseabilidad social de Marlowe y Crowne en una población de adultos. *Salud Mental*, 13, 35-37.
- Lara-Cantú, M.A. y Suzan-Reed, M. (1988). La escala de deseabilidad social de Marlowe y Crowne: un estudio psicométrico. *Salud Mental*, 11, 25-29.
- McDonald, R.P. (1967) *Non linear factor analysis*. Psychometric Monograph n. 15
- McDonald, R.P. y Mok, M.C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, 30, 23-40.
- McDonald, R.P. y Ahlawat, K.S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, 27, 82-99.
- Meehl, P.E. y Hataway, S.R. (1946). The K-factor as a suppressor variable in the MMPI. *Journal of Applied Psychology*, 30, 525-564.
- Michaelis, W. y Eysenck, H.J. (1971). The determination of personality inventory factor patterns and intercorrelations by changes in real-life motivation. *Journal of Genetical Psychology*, 118, 223-234.
- Millham, J. (1974). Two components of need for approval score and their relationship to cheating following success and failure. *Journal of Research in Personality*, 8, 378-392.
- Mislevy, R.J. y Bock, R.D. (1990). *BILOG 3. Item analysis and test scoring with binary logistic models*. Mooresville: Scientific Software.
- Muñiz, J. (1990). *Teoría de respuesta a los ítems*. Madrid: Pirámide
- Nunnally, J.C. (1987). *Teoría psicométrica*. México: Trillas.
- O'Grady, K.E. (1988). The Marlowe-Crowne and the Edwards social desirability scales: a psychometric perspective. *Multivariate Behavioral Research*, 23, 87-101.
- Ones, D.S.; Viswesvaran, C. y Reiss, A.D. (1996). Role of social desirability in personality testing for personnel selection: the red herring. *Journal of Applied Psychology*, 81, 660-679.
- Paulhus, D.L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46, 598-609.
- Ramanaiah, N. y Martin, H.J. (1980). On the two-dimensional nature of the Marlowe-Crowne social desirability scale. *Journal of Personality assessment*, 44, 507-514.
- Reise, S.P. (1999). Personality measurement issues viewed through the eyes of IRT. En: S.E. Embretson y S.L. Hershberger. (eds.) *The new rules of measurement* (pp. 219-241). Hillsdale: LEA.
- Reise, S.P. y Waller, N.G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement*, 14, 45-58.
- Reynolds, W.M. (1992). Development of reliable and valid short forms of the Marlowe-Crowne social desirability scale. *Journal of Clinical Psychology*, 38, 119-125.
- Rorer, L.G. (1965). The great response-style myth. *Psychological Bulletin*, 63, 129-156.
- Schmitt, M.J. y Steyer, R. (1993). A latent state-trait model (not only) for social desirability. *Personality and Individual Differences*, 14, 519-529.
- Shultz, K.S. y Chávez, D.V. (1994). The reliability and factor structure of a social desirability scale in english and in spanish. *Educational and Psychological Measurement*, 54, 935-940.
- Steinberg, L. y Thissen, D. (1995). Item response theory in personality research. En: P.E. Shrout y S.T. Fiske (eds.) *Personality research methods, and theory* (pp. 161-181). Hillsdale: LEA.
- Strahan, R. y Gerbasi, K.C. (1972). Short, homogeneous versions of the Marlowe-Crowne social desirability scale. *Journal of Clinical Psychology*, 28, 191-193.
- Tanaka, J.S. y 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.
- Viswesvaran, C. y Ones, D.S. (1999). Meta-analyses of fakability estimates: implications for personality measurement. *Educational and Psychological Measurement*, 59, 197-210.
- Waller, N.G.; Tellegen, A.; McDonald, R.P. y Lykken, D.T. (1996). Exploring nonlinear models in personality assessment: development and validation of a negative emotionality scale. *Journal of Personality*, 64, 545-576.