Psicothema 2005. Vol. 17, nº 3, pp. 484-491
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Propiedades psicométricas de un test adaptativo informatizado para la medición del ajuste emocional

David Aguado*, Víctor J. Rubio, Pedro M. Hontangas** y José Manuel Hernández
Universidad Autónoma de Madrid, * Instituto de Ingeniería del Conocimiento y ** Universidad de Valencia

En el presente trabajo se describen las propiedades psicométricas de un Test Adaptativo Informatizado para la medición del ajuste emocional de las personas. La revisión de la literatura acerca de la aplicación de los modelos de la teoría de la respuesta a los ítems (TRI) muestra que ésta se ha utilizado más en el trabajo con variables aptitudinales que para la medición de variables de personalidad, sin embargo diversos estudios han mostrado la eficacia de la TRI para la descripción psicométrica de dichas variables. Aun así, pocos trabajos han explorado las características de un Test Adaptativo Informatizado, basado en la TRI, para la medición de una variable de personalidad como es el ajuste emocional. Nuestros resultados muestran la eficiencia del TAI para la evaluación del ajuste emocional, proporcionando una medición válida y precisa, utilizando menor número de elementos de medida en comparación con las escalas de ajuste emocional de instrumentos fuertemente implantados.

Psychometric properties of an emotional adjustment computerized adaptive test. In the present work it was described the psychometric properties of an emotional adjustment computerized adaptive test. An examination of Item Response Theory (IRT) research literature indicates that IRT has been mainly used for assessing achievements and ability rather than personality factors. Nevertheless last years have shown several studies wich have successfully used IRT to personality assessment instruments. Even so, a few amount of works has inquired the computerized adaptative test features, based on IRT, for the measurement of a personality traits as it's the emotional adjustment. Our results show the CAT efficiency for the emotional adjustment assessment so this provides a valid and accurate measurement; by using a less number of items in comparison with the emotional adjustment scales from the most strongly established questionnaires.

El ajuste emocional es uno de los constructos clave en la determinación de la «estructura de la personalidad» desde los modelos del rasgo. Puede observarse su presencia en las teorías de Guilford (Guilford, 1959), de Eysenck (Eysenck, 1947) o de Cattell (Cattell y Scheier, 1961). Y actualmente, en el modelo de los cinco grandes factores (Five Factor Model, FFM) de personalidad (Digman, 1990; Goldberg, 1990), en el que se incluye la estabilidad emocional como uno de los «Big Five». Queda definida como el grado en el que un individuo es inseguro, ansioso y deprimido emocional frente a calmado, confiado y frío. De acuerdo con el NEO-PI-R (Costa y McCrae, 1985, 1992), la estabilidad emocional estaría compuesta por seis facetas: ansiedad, hostilidad, depresión, ansiedad social, impulsividad y vulnerabilidad, todas ellas incluidas en una sola dimensión. Igualmente, la dimensión de ajuste emocional es considerada en los más habituales cuestionarios de personalidad: el 16 PF de Cattell (Cattell, 1972), el EPQ de Eysenck (Eysenck y Eysenck, 1978), el BFQ (Caprara, Barbanelli y Borgogni, 1993) y el mencionado anteriormente NEO-PI.

Fecha recepción: 5-7-04 • Fecha aceptación: 23-12-04
Correspondencia: David Aguado García
Instituto de Ingeniería del Conocimiento
Universidad Autónoma de Madrid
28049 Madrid (Spain)
E-mail: david.aguado@iic.uam.es

Por otro lado, la teoría de la respuesta al ítem (TRI) proporciona un marco preciso y con importantes ventajas sobre la teoría clásica de tests (TCT), para el análisis de las propiedades psicométricas de una medida psicológica, como han expresado entre otros Lord (1952, 1953, 1980), Birnbaum (1968), Hambleton y Swaminathan (1985), Hambleton (1990), Hambleton, Swaminathan y Rogers (1991), y Van der Linden y Hambleton (1997). Esta aproximación psicométrica en conjunción con los avances en el uso generalizado de los ordenadores personales hace posible lo que se vienen denominando Tests Adaptativos Informatizados (TAIs) (Lord y Novick, 1968; Lord, 1970; Owen, 1975). Básicamente un TAI es una prueba de evaluación psicológica que se administra a través del ordenador y cuya característica distintiva es que la presentación de los ítems se realiza en función del nivel en la variable medida que va demostrando el evaluado (Olea y Ponsoda, 1996). Una revisión general sobre la literatura psicométrica muestra que la TRI ha sido utilizada fundamentalmente para la evaluación de las capacidades y no tanto para la medición de factores de personalidad (Reise y Henson, 2003), sin embargo, cada vez un mayor número de trabajos muestran la viabilidad de la aplicación de los presupuestos de la TRI a variables de personalidad (Reise y Waller, 1990; Ferrando, 1994, 2001; Flannery, Reise y Widaman, 1995; Cooke y Michie, 1997; Zumbo, Pope, Watson y Hubley, 1997; Gray-Little, Williams y Hankock, 1997; Rouse, Finger y Butcher, 1999; Aguado, Santa Cruz, Dorronsoro y Rubio, 2000;

Robie, Zickar y Schmit, 2001; Chernyshenko, Stark, Chan, Drasgow y Williams, 2001), utilizando diferentes modelos según el tipo de ítems y el número de parámetros. En el caso de ítems dicotómicos, algunos estudios muestran que el modelo de dos parámetros ajusta generalmente bien a los datos (Ferrando, 1994, 2001; Reise y Waller, 1990), mientras que otros sugieren la utilización de modelos de tres parámetros (Zumbo et al., 1997; Rouse et al., 1999) incluyendo el parámetro de pseudo-azar como un acercamiento al control de la deseabilidad social (Rouse et al., 1999). En el caso de ítems politómicos, el modelo de respuesta graduada de Samejima (1969) (Graded Response Model, GRM) ha mostrado un buen funcionamiento en la mayoría de los trabajos (Flannery et al., 1995; Cooke y Michie, 1997; Gray-Little et al., 1997; Robie et al., 2001), a excepción de Chernyshenko et al. (2001).

Teniendo en cuenta los anteriores trabajos, que han descrito el funcionamiento de escalas de personalidad desde la TRI, se observa la ausencia de estudios sobre la posibilidad de administrar estas pruebas de forma adaptativa (Reise y Henson, 2000). Por ello, el objetivo del presente trabajo es analizar la viabilidad de un test adaptativo informatizado para la medición del ajuste emocional y describir sus propiedades psicométricas. Para ello se ha utilizado el GRM de Samejima (1969). Aunque existen modelos alternativos basados en diferentes planteamientos teóricos (van der Linden y Hambleton, 1997) que podrían ser aplicables, como el Modelo de Crédito Parcial de Masters (1982) (Partial Credit Model, PCM) y sus variantes, utilizamos el GRM por las siguientes razones: a) es uno de los primeros modelos desarrollados para ítems politómicos graduados; por tanto, es bien conocido y tiene una amplia tradición; b) es adecuado para ítems politómicos graduados con diferentes parametros ‘a’ frente a modelos que consideran que el parámetro ‘a’ es igual para todos, como el PCM; c) el GRM ha sido más utilizado con escalas de respuesta tipo Likert (escala habitual para la medida de características de personalidad), mientras que el PCM se ha utilizado más para considerar las etapas superadas en la resolución de una tarea o situaciones que reflejan una puntuación parcial; d) existen más estudios publicados sobre recuperación de parámetros frente al PCM, por lo que se conocen bien las condiciones para obtener estimaciones adecuadas; y e) en los trabajos que comparan el GRM y el PCM con ítems graduados se encuentra un mejor ajuste con el primero (King, King, Fairbank, Schlengar y Surface , 1993; Baker, Rounds y Zevon, 2000).

Método

Participantes

La muestra utilizada en el estudio está formada por 858 estudiantes de Psicología de la Universidad Autónoma de Madrid, de los cuales el 80.6% son mujeres y el 19.4% hombres, con un rango de edad entre 18 y 55 años (media= 19.94, mediana= 19, moda= 18).

Materiales

Escala de ajuste emocional «EAE» (véase Tabla 1). Incorporada como parte de un cuestionario de personalidad desarrollado a medida para fines de selección de personal (CPSP). La escala está compuesta por 28 ítems con seis opciones de respuesta en un gradiente de acuerdo-desacuerdo.

Escala de ajuste emocional presente en el cuestionario Big Five Questionnaire (BFQ) (Caprara, Barbanelli y Borgogni, 1993). La escala está compuesta por 23 ítems de 5 opciones de respuesta.

Escala de Neuroticismo presente en el Eysenck Personality Questionnaire (EPQ-A) (Eysenck y Eysenck, 1978). La escala está compuesta por 25 ítems de verdadero-falso.

Procedimiento de recogida de datos

EL CPSP fue administrado en su formato íntegro a 440 estudiantes. Tres meses después a otros 418 estudiantes se les administró únicamente la escala de ajuste emocional EAE, junto con la escala de ajuste emocional del BFQ y la escala de neuroticismo del EPQ. Se completó así la muestra de 858 participantes. Todos los instrumentos fueron administrados en su versión de lápiz y papel.

Análisis de datos

Para el establecimiento de la unidimensionalidad de la EAE se realizaron análisis factoriales encaminados a observar la existencia de una dimensión dominante. La bondad de ajuste del modelo fue deducida de la adecuación del proceso de estimación y de la invarianza de parámetros. Normalmente, el ajuste de este modelo suele comprobarse mediante varios indicadores indirectos, dado que no existen tests de bondad de ajuste comúnmente aceptados para el modelo de respuesta graduada (salvo en el caso de comparación de modelos anidados), entre los que se encuentran: la convergencia en el proceso de estimación del modelo, la obtención de valores razonables para los parámetros, la existencia de errores típicos de estimación pequeños y la aportación de evidencia sobre la invarianza de parámetros de ítems y sujetos. Para la determinación de las propiedades de la EAE, se estudió: a) la validez a partir de la estructura factorial de la EAE y de las correlaciones con las otras dos escalas criterio; b) la precisión, a partir del error típico de estimación (Se), el error absoluto medio (EAM, diferencias absolutas entre datos reales y estimaciones) y el sesgo (diferencias algebraicas entre datos reales y estimaciones). Igualmente, para el establecimiento de las propiedades del TAI, se estudiaron las mismas características anteriores y la eficiencia del procedimiento (número de ítems necesarios para alcanzar cierta precisión) con diferentes criterios de parada (error típico de estimación inferior a .30, .35 y .40). Todos los análisis se realizaron con el programa SPSS 11.5, excepto la calibración, que se realizó mediante Parscale 3.0 (Muraki y Bock, 1996), y las simulaciones, que se ejecutaron mediante programación en C++ desarrollada al efecto.

Procedimiento de simulación

Se realizaron dos tipos de simulaciones. El procedimiento seguido en ambas es básicamente el mismo, variando únicamente los sujetos utilizados y la forma en que se emiten las respuestas. En la Simulación I, las respuestas son generadas teóricamente bajo criterios probabilísticos, mientras que en la Simulación II, las respuestas son obtenidas del cuestionario de lápiz y papel.

En la Simulación I se generan 100 sujetos para 12 niveles de rasgo comprendidos entre –2.5 y +2.5. En total 1.200 sujetos. Dichos valores son el nivel de rasgo real de cada sujeto simulado. La emisión de la respuesta se realiza mediante criterios probabilísticos. Una vez el programa presenta un ítem, se genera un número aleatorio entre 0 y 1, y se compara dicho número con las probabi-

lidades de respuesta a las diferentes categorías del ítem en función del Modelo de Respuesta Graduada. Se asigna la respuesta asociada al rango de probabilidad en el que se encuentre el número aleatorio. De modo que, para un ítem con las siguientes probabilidades de respuesta: P_{Cat1}= 0.33, P_{Cat2}= 0.41, P_{Cat3}= 0.18, P_{Cat4}= 0.05, P_{Cat5}= 0.02 y P_{Cat6}= 0.01, se calculan las probabilidades acumuladas: PA_{Cat1}= 0.33, PA_{Cat2}= 0.74, PA_{Cat3}= 0.82, PA_{Cat4}= 0.87, PA_{Cat5}= 0.89 y PA_{Cat6}= 1, y si el número aleatorio es inferior o igual a 0.33, la respuesta es la categoría 1; si el número aleatorio está entre 0.33 y 0.74, se le asigna la categoría 2, y así sucesivamente. La lógica de la simulación estriba en que las probabilidades asociadas a las categorías estarán en función del nivel de rasgo del sujeto, de este modo un sujeto con muy poco nivel de rasgo tendrá asociada una probabilidad muy alta de responder a la categoría 1. Así, al extraer un número aleatorio entre 0 y 1 lo más probable es que salga entre el rango de probabilidad cubierto por las categorías bajas.

En la Simulación II se trabaja directamente con las respuestas reales que los sujetos emitieron al cuestionario en lápiz y papel. Las respuestas se encuentran almacenadas en un fichero, de modo que, cuando el ordenador decide el ítem que va a ser administrado, una función se encarga de buscar en el fichero de datos la respuesta que en su momento el sujeto emitió ante ese ítem.

El TAI en ambos tipos de simulaciones tiene las siguientes características. La sesión se inicia asignando al sujeto una θ aleatoria entre –1 y +1. Para la θ asignada, el programa busca el ítem más informativo según Samejima (1997), emite la respuesta según los criterios de simulación especificados anteriormente y estima el nivel de rasgo y el error de medida mediante el procedimiento de máxima verosimilitud. Después, se comprueba si se cumple el criterio de parada, si así es, se da por finalizada la sesión, y si no se cumple dicho criterio, se vuelve a buscar el ítem máximamente informativo para la nueva θ estimada. Si el sujeto en el primer ítem responde en la categoría más alta o más baja, no es posible aplicar máxima verosimilitud, y, en este caso, se aplica el procedimiento stepsize variable (Dodd, De Ayala y Koch, 1995) para la estimación del siguiente.

Resultados

Unidimensionalidad

La unidimensionalidad hace referencia a que la probabilidad de las respuestas de los sujetos a los ítems depende única y exclusivamente de un rasgo. En la práctica el requerimiento teórico de la unidimensionalidad absoluta raramente se cumple. Una aproximación más realista consiste en demostrar que existe una dimensión dominante (Hambleton y Swaminathan, 1985; Hambleton, Swaminathan y Rogers, 1991). En este sentido, el porcentaje de varianza explicada por el primer factor respecto del explicado por el segundo puede ser computado a través del cociente entre los autovalores de ambos factores. En nuestro caso dicho cociente es igual a 5.43 (Autovalor Factor 1= 8.86; Autovalor Factor 2= 1.63), lo que indica la unidimensionalidad de la escala (Martínez Arias, 1995). Además, el primer factor explica un porcentaje de varianza cercano al 32%. Algunos autores establecen que por encima del 20% se podría hablar de unidimensionalidad (Reckase, 1979), sin embargo otros autores apuntan hacia el 40% (Carmines y Zeller, 1979). Por otro lado, siguiendo la estrategia de Horn (1965), que consiste en comparar el gráfico de sedimentación de la solución factorial de la escala y la de una matriz de datos aleatorios con las mismas características se observa cierta unidimensionalidad (Gráfico 1). Igualmente, las saturaciones en el primer factor (Tabla 1) también parecen apuntar hacia la existencia de un factor dominante, ya que todos los ítems presentan altas saturaciones. Con dichas evidencias podríamos considerar que la EAE es unidimensional a efectos de su utilización desde la TRI.

Estimación de parámetros y ajuste del modelo

La calibración de la EAE fue realizada utilizando el Modelo de Respuesta Graduada de Samejima (1969). Aplicando el modelo, cada ítem se caracteriza por un parámetro de discriminación (a_i) y cinco parámetros de localización de las categorías de respuesta (b_{ik}), es decir, uno menos que el número de alternativas de respuesta. La estimación fue realizada en 15 iteraciones, por lo que no hubo problemas de convergencia, y los parámetros y sus errores de estimación (Se) presentan valores razonables. Como se observa en la Tabla 1, los errores de estimación de los parámetros son relativamente pequeños, los valores de a_i están comprendidos entre Se_{item6}= .01 y Se_{item18}= .04; y los de b_{ik} están comprendidos entre Se = .05 y $Se_{item16,K=5}$= .28.

Atendiendo a la discriminación de los ítems, se puede observar que, de acuerdo con la clasificación de Baker (1985), no existen ítems con muy bajo nivel de discriminación y solo uno presenta un índice bajo (ítem 6). La correlación entre los parámetros a_i y la correlación ítem-total corregida, como una medida de la capacidad de discriminación del ítem desde la TCT es .89 (p<.01). Prestando atención a los parámetros b_{ik}, especialmente los extremos b_{i1} y b_{i5}, que representan el nivel de rasgo para el que el ítem ofrece la discriminación, se observa cómo se cubre casi todo el continuo del nivel de rasgo.

Finalmente, la correlación entre las puntuaciones estimadas con los ítems pares y con los ítems impares de la EAE resultó ser .70 (p<.01); y las correlaciones entre los parámetros de los ítems estimados con una mitad de los sujetos y con la otra mitad de los sujetos fueron para el parámetro a_i .81 (p<.01), y para los parámetros b_{ik}: b_{i1}= .87; b_{i2}= .84; b_{i3}= .82; b_{i4}= .77; b_{i5}= .75; todas significativas estadísticamente (p<.01). Dichas correlaciones apun-

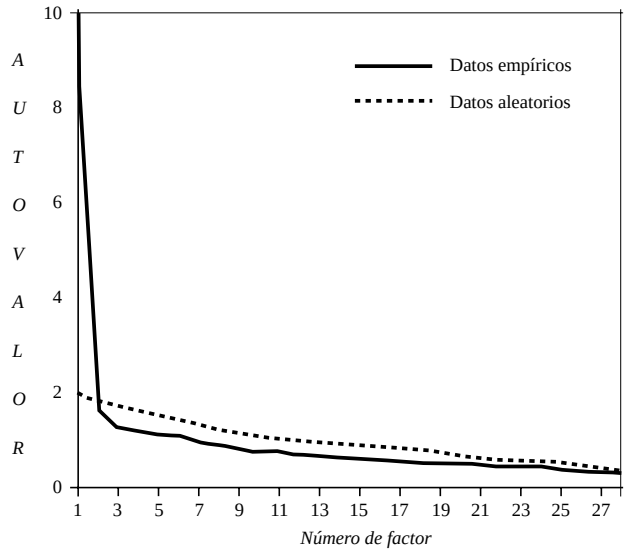

Gráfico 1. *Diagrama de sedimentación de los datos empíricos y aleatorios*

tan hacia el cumplimiento moderado de la invarianza de los parámetros. En resumen, los resultados sugieren un ajuste razonable del modelo de respuesta graduada, que permite que la EAE pueda ser administrada bajo un formato de test adaptativo informatizado.

Propiedades del banco de ítems

Validez. Del análisis factorial realizado sobre los 28 ítems de la EAE se desprende una estructura de componentes muy similar a la presentada teóricamente por otros autores (Buss y Plomin, 1975; Guilford, 1959; Eysenck, 1947; Cattell y Scheier, 1961; Costa y McCrae, 1985, 1992). Los cinco componentes extraídos explicaron el 54.4% de la varianza y mediante rotaciones oblicuas y ortogonales se reprodujeron las mismas asociaciones de ítems, que etiquetamos como ansiedad, hostilidad, desconfianza en sí mismo, inestabilidad del ánimo, y tendencias depresivas. En el análisis factorial de segundo orden llevado a cabo sobre las puntuaciones factoriales obtenidas en el primer análisis factorial, emergió un único factor indicando una dimensión dominante subyacente a la estructura factorial presentada anteriormente.

Respecto de la validez convergente del banco se observa cómo

Tabla 1
Descriptivos, parámetros estimados y error típico de estimación (Se) de la EAE

Item	***Me.***	S_x	r_{it}	F_1	a_i (S_e)	b_i (S_e)	b_2 (S_e)	b_3 (S_e)	b_4 (S_e)	b_5 (S_e)
1. Mis nervios llegan a tal límite que ciertos sonidos (por ejemplo, el chirrido de una puerta) se me hacen insoportables	4.24	1.42	.50	.54	.68(.02)	-3.24(.15)	-1.98(.09)	-.84(.07)	.02(.07)	1.42(.08)
2. Creo que soy más nervioso que la mayoría	3.72	1.48	.53	.58	.70(.02)	-2.64(.12)	-1.29(.07)	-.20(.06)	.77(.07)	1.83(.09)
3. Mis músculos suelen estar en tensión	3.92	1.34	.53	.57	.73(.02)	-3.02(.14)	-1.67(.08)	-.48(.06)	.45(.06)	2.02(.09)
4. Noto palpitaciones o golpes en el corazón	3.79	1.55	.48	.53	.65(.02)	-2.62(.12)	-1.40(.08)	-.26(.07)	.54(.07)	1.77(.09)
5. En situaciones de presión tengo trastornos digestivos	3.68	1.67	.41	.45	.50(.02)	-2.67(.13)	-1.08(.09)	-.10(.08)	.76(.09)	1.89(.11)
6. No he llevado un tipo de vida ordenado y normal	4.43	1.43	.31	.35	.38(.01)	-5.23(.27)	-3.24(.16)	-1.68(.12)	-.45(.11)	1.61(.12)
7. Algunas veces estoy de tan mal humor que me dan ganas de ponerme a romper cosas	3.66	1.67	.54	.59	.82(.03)	-1.84(.08)	-.78(.06)	-.11(.06)	.46(.06)	1.48(.07)
8. Las contrariedades muy pequeñas me irritan mucho	3.79	1.34	.62	.67	.98(.03)	-2.47(.10)	-1.35(.06)	-.27(.05)	.62(.05)	1.80(.07)
9. Me sacan de quicio de un modo insoportable pequeñas cosas, aunque reconozca que son triviales	3.42	1.45	.61	.66	.94(.03)	-1.96(.08)	-.78(.05)	.16(.05)	.87(.06)	2.09(.09)
10. A menudo, cuando algo me altera, pierdo la cabeza y hago tonterías	4.29	1.33	.44	.49	.60(.02)	-4.26(.24)	-2.30(.11)	-1.09(.08)	.07(.07)	1.62(.09)
11. Suelo enfadarme con las personas demasiado pronto	4.07	1.38	.55	.59	.79(.03)	-2.97(.14)	-1.70(.08)	-.64(.06)	.25(.06)	1.62(.08)
12. Las luchas más encarnizadas las tengo conmigo mismo	3.15	1.54	.50	.55	.68(.02)	-1.70(.09)	-.53(.07)	.44(.07)	.33(.08)	2.59(.12)
13. Tengo recuerdos o pensamientos que me dan vueltas constantemente en la cabeza	2.39	1.32	.56	.61	.83(.03)	-.79(.06)	.37(.06)	1.28(.07)	.14(.09)	3.25(.16)
14. Todavía me preocupan seriamente errores que cometí en el pasado	3.34	1.56	.53	.58	.74(.03)	-1.90(.09)	-.69(.06)	.29(.06)	.95(.07)	2.12(.09)
15. Me cuesta bastante concentrarme en una tarea o trabajo	3.75	1.34	.41	.45	.51(.02)	-3.98(.20)	-1.89(.10)	-.25(.08)	.95(.09)	2.86(.14)
16. Se me ocurre lo que debería haber dicho y hecho cuando ya ha pasado el momento	2.62	1.25	.44	.48	.56(.02)	-1.69(.09)	.04(.08)	1.52(.09)	.87(.13)	4.69(.28)
17. Mis emociones son tan poco lógicas que soy incapaz de controlarlas	4.17	1.34	.57	.62	.84(.03)	-2.93(.14)	-1.73(.08)	.80(.06)	.14(.06)	1.49(.07)
18. Me siento unas veces triste y otras alegre sin motivo	3.26	1.55	.63	.68	1.02(.04)	-1.42(.06)	-.57(.05)	.27(.05)	.90(.05)	1.92(.08)
19. A menudo me siento cansado e indiferente sin ninguna razón para ello	3.31	1.41	.60	.65	.93(.03)	-1.93(.08)	-.66(.05)	.26(.05)	1.08(.06)	2.20(.09)
20. Tengo sentimientos de desasosiego como si deseara algo pero sin saber qué	3.33	1.46	.63	.68	.99(.03)	-1.66(.07)	-.73(.05)	.15(.05)	1.00(.06)	2.02(.08)
21. Suelo tener un estado de ánimo bastante estable	3.56	1.49	.61	.65	.93(.03)	-1.83(.08)	-.82(.06)	-.17(.05)	.68(.05)	2.02(.08)
22. Lloro con facilidad	3.35	1.68	.35	.39	.47(.02)	-2.20(.12)	-.74(.09)	.28(.09)	1.18(.10)	2.60(.13)
23. Cuando me desanimo me cuesta recuperarme	3.69	1.35	.56	.60	.78(.03)	-2.67(.12)	-1.36(.07)	-.21(.06)	.77(.06)	2.16(.09)
24. Me gustaría ser tan feliz como los demás	3.73	1.58	.51	.56	.69(.02)	-2.27(.10)	-1.25(.07)	-.26(.07)	.63(.07)	1.73(.09)
25. Me considero una persona feliz y contenta	4.39	1.23	.44	.45	.59(.02)	-4.04(.21)	-2.81(.13)	-1.68(.09)	-.17(.07)	1.76(.09)
26. Me pongo nervioso cuando pienso en todas las cosas que tengo que hacer	2.84	1.31	.49	.54	.69(.02)	-1.74(.09)	-.28(.07)	1.02(.07)	1.97(.09)	3.31(.16)
27. Tengo muy poca confianza en mí mismo	3.83	1.53	.56	.60	.73(.03)	-2.34(.10)	-1.27(.07)	-.37(.06)	.42(.06)	1.79(.08)
28. No me desanimo ante las dificultades cotidianas	4.06	1.31	.37	.41	.51(.02)	-4.35(.23)	-2.46(.12)	-1.01(.09)	.37(.09)	2.56(.12)
TOTAL	101.7	22.4	α= .92							

Me.= Media; S_x= desviación típica; r_{it}= correlación ítem-total corregida; F_1= saturaciones en el primer factor.

las correlaciones obtenidas entre las puntuaciones en nuestra EAE y las puntuaciones obtenidas en la escala N del EPQ y la escala de ajuste emocional del BFQ son altas y significativas (véase tabla 2). Es necesario resaltar que la más alta de las correlaciones es la producida entre nuestra escala y la escala N del EPQ (.893, p<.01), superior incluso a la encontrada entre la escala de ajuste emocional del BFQ y la escala N del EPQ (.850 , p<.01). La menor de las tres correlaciones es la encontrada entre la escala EAE y la escala de ajuste emocional del BFQ (.848, p < .01). Dichas correlaciones descienden en su cuantía cuando la puntuación en la escala es estimada mediante los procedimientos de máxima verosimilitud utilizados en la TRI. La máxima correlación en este caso es entre el BANCO y la puntuación obtenida en la EAE (.776, p<.01). Igualmente la correlación con la puntuación en BFQ y en EPQ es también significativa y con unos valores similares.

En conjunto, tanto el análisis de correlaciones como los análisis factoriales realizados, parecen apuntar hacia la validez de la EAE en términos de convergencia de puntuaciones y contenido.

Precisión. En el Gráfico 2 se presenta la precisión de la EAE y la comparación del error típico de estimación de la TRI con el de la TCT a partir de la fiabilidad de la escala (*alpha*= .92). Como se puede observar, la escala no tiene el mismo error de medida para los diferentes niveles de rasgo. Es especialmente precisa para los niveles medios. En el rango acotado por $-2 < \theta < +2$, en el cual, teniendo en cuenta la normalidad del continuo del rasgo, se acumulan el 96% de los sujetos, la escala presenta un error típico de medida inferior a .27. Esto significa una precisión buena para un altísimo porcentaje de la población general. El menor error de medida se comete, no obstante, sobre un rango menor del continuo del nivel de rasgo, el comprendido entre $-1 < \theta < +1$, donde el error de medida se aproxima a .24. Por otro lado, los mayores errores de medida se cometen al tratar de estimar los niveles de rasgo correspondientes a $\theta = -3$, $\theta = -4$, $\theta = +3$, $\theta = +4$, en los que el error de medida aumenta hasta .5. Además, son inferiores sensiblemente los errores extremos en el polo negativo frente al polo positivo. Por tanto, la escala muestra suficiente precisión para la evaluación de personas con niveles bajos, medios-bajos, medios, medios-altos y altos de ajuste emocional, pero no se muestra tan precisa al evaluar a sujetos con niveles extremos (ya sean positivos o negativos). Así, podemos observar cómo, comparado con el supuesto de igualdad de los errores de estimación a lo largo del continuo del nivel de rasgo (TCT), en realidad, se están infraestimando (midiendo con mayor precisión de la informada) los niveles medios, y sobreestimando (midiendo con menor precisión de la informada) los niveles extremos. Por otro lado, las estimaciones son insesgadas, como se muestra en la tabla 3. La media de las diferencias algebraicas encontradas entre la θ real y la θ estimada en la *Simulación I*, utilizando todo el banco de ítems, resultó ser cero. Teniendo en cuenta los diferentes niveles de rasgo, en ningún caso dicha diferencia llega a estar una décima por encima o por debajo de cero. Ello parece indicar que mediante el procedimiento de estimación de θ no se están produciendo estimaciones sistemáticamente superiores o inferiores en ningún punto del continuo del nivel de rasgo, respecto del nivel real de los sujetos. En definitiva, desde la TRI se obtiene la precisión con la que se ha estimado la puntuación de un evaluado. Este aspecto unido a la invarianza de los parámetros presentada anteriormente, por la que una persona recibe la misma puntuación aunque se le presenten ítems distintos, justifica el uso de los tests adaptativos.

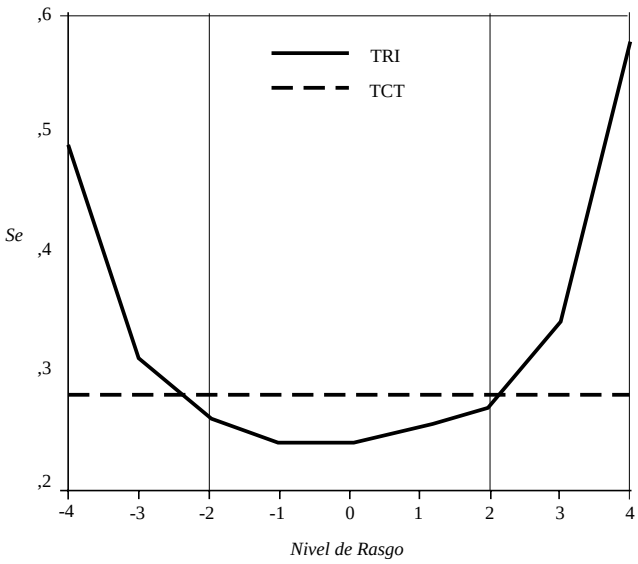

Gráfico 2. *Precisión del banco de ítems según la TCT y la TRI (S_e, error típico de medida)*

Tabla 2
Correlaciones

	Correlaciones					
	EPQ	**EAE**	**BANCO**	**TAI-30**	**TAI-35**	**TAI-40**
BFQ	.850**	.848**	.666**	.614**	.595**	.575**
EPQ		.893**	.697**	.664**	.614**	.592**
EAE			.776**	.745**	.697**	.676**
BANCO				.976**	.919**	.884**
TAI-30					.944**	.908**
TAI-35						.970**

BFQ= total en escala de ajuste emocional del BFQ; EPQ= total en escala N del EPQ-A; EAE= total en la escala EAE; BANCO= θ con todos los ítems; TAI-30= θ en el TAI con criterio de parada Se <.30; TAI-35= θ TAI con criterio de parada Se <.35; TAI-40= θ TAI con criterio de parada Se <.40.
***Correlación significativa a nivel p<.01*

Propiedades del TAI

Validez. Como se observa en la tabla 2, las correlaciones entre las puntuaciones criterio (BFQ, EPQ, EAE y BANCO) y las puntuaciones estimadas en las diferentes condiciones TAI son significativas en todos los casos y de una cuantía importante (con el BANCO) y moderada (con el resto de criterios para los que la puntuación se estima mediante procedimientos de la TCT). Por otro lado, es necesario señalar cómo la cuantía de las correlaciones desciende a medida que se relaja el criterio de precisión del TAI. Así, la mayor correlación con los criterios aparece siempre en la condición en la que el TAI tiene un criterio de parada en Se<.30. En dicha condición la correlación con el BANCO es la mayor (.976, p<.01), y se muestran correlaciones moderadas con BFQ (.614, p<.01), con EPQ (.664, p<.01), y con EAE (.745, p<.01). En resumen, se observa cómo el TAI en la condición de uso más habitual (TAI-30), mantiene razonablemente bien las propiedades de validez estimadas para el banco en su conjunto. Además, con un criterio de parada más laxo (TAI-40), dichas propiedades se ven afectadas pero mantienen unos niveles aceptables.

Precisión. En la *Simulación I*, las diferencias medias en valor absoluto entre la θ estimada y la θ real asignada, para los diferentes criterios de parada del TAI resultaron ser de .23 (Se < .30), .29 (Se < .35), .34 (Se < .40). Con ello se observa cómo las propiedades originales del bando de ítems se van deteriorando en el TAI a medida que se es menos exigente con el criterio de parada establecido. Por otro lado, los anteriores valores se mantienen relativamente estables a lo largo del continuo del nivel de rasgo. En cuanto al sesgo, se puede observar en la tabla 3 cómo se reproducen los efectos presentados para el banco de ítems. En ningún punto del continuo del nivel de rasgo, ni en la media del continuo, parece existir una tendencia a que se estimen sistemáticamente puntuaciones superiores o inferiores a las reales.

Eficiencia. Respecto de la eficiencia del TAI, entendiendo ésta como el número de ítems necesarios para obtener una determinada precisión, comprobamos cómo en la Simulación I (tabla 3) a medida que se utilizan más ítems y que el error de medida exigido es más pequeño, la diferencia entre la puntuación estimada y la puntuación real del sujeto simulado es menor. Con un criterio de parada establecido en Se < .30 se utilizan como media 18.8 ítems, y para los niveles centrales del rasgo ($-1 < \theta < +1$) se utilizan 16 ítems. Siendo un poco menos exigentes en la precisión y estableciendo el criterio de parada en Se < .35, en los niveles centrales del rasgo son necesarios únicamente 9 ítems, y en todo el continuo una media cercana a los 10 ítems. Estos resultados muestran claramente cómo manteniendo el nivel de precisión requerido para la estimación del ajuste emocional de las personas, mediante un TAI se puede ser más eficiente en la evaluación en términos de un menor número de ítems utilizados.

Discusión

En primer lugar, es necesario señalar que tanto la unidimensionalidad, establecida mediante diversos procedimientos, como el ajuste del GRM de Samejima a los datos, en línea con los hallazgos de los estudios precedentes (Reise y Waller, 1990; Ferrando, 1994, 2001; Flannery et.al., 1995; Cooke y Michie, 1997; Zumbo et al., 1997; Gray-Little et. al., 1997; Rouse et. al., 1999; Robie et. al., 2001; Chernyshenko et. al., 2001), permiten la administración de la EAE en formato adaptativo. En este sentido, el TAI ofrece una adecuada precisión. Tanto el error de medida estimado como las diferencias entre nivel de rasgo real y nivel estimado mediante el TAI en las diferentes simulaciones así lo muestran. Respecto a los errores de medida, al utilizar todo el banco de ítems Se es próximo a .24 en el rango establecido en $-1 < \theta < +1$, y se aproxima a .27 en el rango $-2 < \theta < +2$, sensiblemente inferiores a los esta-

Tabla 3
Resultados de la simulación I

θ	Banco		Criterios de parada del TAI								
			Se <.30			Se <.35			Se <.40		
	Sesgo	EAM	Sesgo	EAM	*N*	Sesgo	EAM	*N*	Sesgo	EAM	*N*
-2.50	.01	.21	.02	.24	23	.08	.33	11.3	.07	.34	8.5
-2.00	-.01	.17	-.02	.19	19	-.01	.31	9.9	.00	.3	7.4
-1.50	.03	.20	.01	.21	17.3	.00	.29	9.2	-.03	.31	7.2
-1.00	.02	.18	.05	.24	16.5	-.04	.26	9	-.01	.36	7
-.50	.00	.19	-.05	.21	16.1	.03	.32	9	-.03	.37	7
-.25	-.02	.21	.01	.24	16.2	.05	.27	9	.01	.3	7
.25	-.02	.20	.03	.25	16.3	-.01	.3	9	.00	.34	7
.50	.02	.20	-.00	.25	16.3	.08	.3	9	-.01	.34	7
1.00	.00	.18	-.03	.25	17	-.08	.3	9	.04	.34	7
1.50	-.01	.23	-.01	.23	18.4	-.04	.29	9.5	.03	.3	7.3
2.00	.00	.24	.05	.23	22.5	-.02	.25	10.6	-.09	.37	7.9
2.50	.03	.24	.08	.27	26.8	.01	.27	14.2	-.10	.35	9.6
TOTAL	.00	.20	.01	.23	18.8	.00	.29	9.9	-.01	.34	7.5

θ= nivel de rasgo; Se= error típico de estimación; Sesgo= media de diferencias algebraicas entre nivel estimado y real; EAM= media de diferencia absolutas entre nivel estimado y real; N= número de ítems administrados.

blecidos para las escalas de ajuste emocional del EPQ-A (entre .38 y .41), del BFQ (.36) y del NEO-PI-R (entre .26 y .28) a partir de sus índices de fiabilidad. Así, para el rango $-2 < \theta < +2$ el error de medida cometido es satisfactoriamente bajo, inferior al del EPQ-A y BFQ y parejo al del NEO-PI-R. Dichos resultados son especialmente relevantes si comparamos el número de ítems necesarios para la obtención de las mencionadas precisiones. Estableciendo un criterio de parada del TAI en Se<.30, se utilizan una media de 18.8 ítems, y si el criterio de parada se establece en 15 elementos se consigue un Se=.31. Es decir, entre 15 y 19 elementos de medida frente a los 48 utilizados en el NEO-PI-R, a los 23 del BFQ y a los 25 del EPQ-A. Esta buena precisión del TAI con un menor número de ítems queda claramente justificada con el procedimiento adaptativo utilizado. Por otro lado, la apreciación de un error de medida diferencial para los diferentes niveles de rasgo en la EAE puede ser trasladado a los anteriores instrumentos: el ajuste emocional estimado con dichos instrumentos para niveles extremos de ajuste emocional se está realizando con un error de medida superior al informado.

Por otro lado, las correlaciones con otros tests (.893 con la escala N del EPQ y .848 con la escala de ajuste emocional del BFQ) indican la validez convergente de la EAE, siendo incluso superior la correlación entre la EAE y la escala N del EPQ, que la encontrada entre la escala de Eysenck y la del BFQ. Estos resultados son más remarcables aún teniendo en cuenta que la correlación informada entre la escala N del EPQ y la escala de ajuste emocional del NEO-PI-R es *r*= .73 (Avía y Sánchez Bernardos, 1995), y que el manual del BFQ presenta una *r*= .66 entre su escala y la escala N del EPQ. Nuestros datos están en línea con estas otras correlaciones, si bien la cuantía de las mismas es ligeramente superior. De igual forma, las correlaciones establecidas entre las puntuaciones en el banco de ítems, las puntuaciones en las diferentes condiciones TAI y las puntuaciones en las escalas de ajuste emocional de EPQ y BFQ muestran la validez convergente del TAI. Además, la exploración realizada sobre la estructura de la EAE muestra la alta sintonía de los componentes con las facetas propuestas por otros autores. En definitiva, existen evidencias aceptables que sugieren la validez de constructo tanto del banco de ítems que es la EAE, como del TAI.

No obstante, es necesario señalar que, a pesar de los prometedores resultados encontrados respecto de las propiedades del TAI, hay que tener en cuenta el carácter preliminar de los mismos, debido al reducido número de elementos que conforman la EAE. La ampliación en el número de ítems del banco permitiría una mayor eficiencia del TAI y una mayor resistencia al deterioro de sus propiedades respecto de las generales del banco, ya que, entre otros aspectos, el reducido número de ítems impide la adecuada representación de los contenidos presentes en la EAE. Por ello, la ampliación de la EAE constituye la principal línea de trabajo a desarrollar. En este sentido, cuatro son los aspectos sobre los que profundizar: a) la generación automática de ítems aplicada a los constructos de personalidad, que como muestra Rojas (2001) es uno de los aspectos centrales en la investigación en tests adaptativos informatizados; b) la aplicación de los diseños de anclaje para la calibración; c) la implementación de mecanismos de control de la exposición de los ítems y del contenido de los mismos garantizando la representatividad de las facetas del ajuste emocional; y d) el análisis de la eficiencia del TAI en contextos reales de aplicación.

Referencias

Aguado, D., Santa Cruz, C., Dorronsoro, J.R.D. y Rubio, V. J. (2000). Algoritmo mixto mínima entropía-máxima verosimilitud para la selección de ítems en un test adaptativo informatizado. *Psicothema, 12*(2), pp. 12-14.

Avia, M.D. y Sánchez Bernardos, M.L. (1995). *Personalidad: aspectos cognitivos y sociales.* Madrid: Pirámide.

Baker, F.B. (1985). *The basics of item response theory.* Porstsmouth, NH: Heinneman.

Baker, J.G., Rounds, J.B. y Zevon, M.A. (2000). A comparison of graded and Rasch partial credit models with subjetive well-being. *Journal of Educational and Behavioral Statitstics, 25,* 253-270.

Birnbaum, A. (1968). Some latent trait models and their use in inferring a examinee's ability. En F.M. Lord y M. Novick (eds.): *Statistical theories of mental test scores.* Reading Mass, Addison-Wesley.

Buss, A.H. y Plomin, R. (1975). *A temperament theory of personality development.* John Wiley and Sons. Traducción al castellano, Madrid: Marova 1980.

Caprara, G.V., Barbaranelli, L. y Borgogni, L. (1993). *BFQ*. Organizzazioni Specialle. Florencia (versión española, Madrid: TEA Ediciones, 1995).

Carmines, E.G. y Zeller, R.A. (1979). *Reliability and validity assessment.* Beverly Hills, CA: Sage Publications.

Cattell, R.B. y Scheier, I.H. (1961). *The meaning and measurement of neuroticism and anxiety.* New York: Ronald Press.

Cattell, R.B. (1972). *Sixteen Personality Factor. 16 Pf.* Illinois: ITPA (versión española, Madrid: TEA Ediciones, 1975).

Chernyshenko, O.S., Stark, S., Chan, K.Y., Drasgow, F. y Williams, B. (2001). Fitting item response theory models to two personality inventories: issues and insights. *Multivariate Behavioral Research, 36,* 523-562.

Cooke, D.J. y Michie, C. (1997). An item response theory analysis of the Hare Psychopath Checklist-revised. *Psychological Assessment 9,* 3-14.

Costa, P.T. y McCrae, R.R. (1985). *The NEO Personality Inventory Manual*. Odessa, FL.: Psychological Assessment Resources.

Costa, P.T. y McCrae, R.R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, F.L.: Psychological Assessment Resources.

Digman, J.M. (1990). Personality structure: emergence of the five factor model. *Annual Review of Psychology 41,* 417-440.

Dodd, B.G., De Ayala, R.J. y Koch, W.R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement 14,* 59-71.

Eysenck, H.J. (1947). *Dimensions of personality.* London: Routledge and Kegan Paul.

Eysenck, H.J. y Eysenck, S.B.G. (1978). *EPQ Cuestionario de Personalidad, Formas A y J.* Madrid: TEA Ediciones.

Ferrando, P.J. (1994). Fitting item response models to the EPI-A Impulsivity scale. *Educational and Psychological Measurement, 54,* 118-127.

Ferrando, P.J. (2001). The measurement of neuroticism using MMQ, MPI, EPI and EPQ items: a psychometric analysis based on item response theory. *Personality and Individual Differences, 30,* 641-656.

Flannery, W.P., Reise, S.P. y Widaman, K.F. (1995). An item response theory of the general and academic scales of the Self-Description Questionnaire II. *Journal of Research in Personality 29,* 168-188.

Goldberg, L.R. (1990). An alternative «description of personality»: the Big-Five factor structure. *Journal of Personality and Social Psychology, 59,* 1.216-1.229.

Gray-Little, B., Williams, V.S.L. y Hancock, T.D. (1997). A item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 23,* 443-451.

Guilford, J.P. (1959). *Personality*. New York: McGraw-Hill.
Hambleton, R.K. (1990). Item response theroy: introduction and bibliography. *Psicothema, 1, 97-107.*
Hambleton, R.K. y Swaminathan, J. (1985). *Item response theory: principles and aplications.* Boston, M.A: Kluwer.
Hambleton, R.K., Swaminathan, J. y Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.
Horn, J.L. (1965). A rationale and technique for estimating the number of factor in factor analysis. *Psychometrika, 30,* 179-185.
King, D.W., King, L.A., Fairbank, J.A., Schlengar, E. y Surface, C.R. (1993). Enhancing the precision of the Mississippi scale for combat-related posttraumatic stress disorder: an application of item response theory. *Psychological Assessment, 5,* 457-471.
Lord, F.M. (1952). A theory of tests scores. *Psychometric Monographs, 7.*
Lord, F.M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517-549.
Lord, F.M. (1970). Some test theory for tailored testing. En W.H. Holtzman (ed.): *Computer asisisted instruction, testing and guidance* (pp. 139-183). New York: Harper and Row.
Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: LEA.
Lord, F.M. y Novick, M.R. (1968). *Statistical theories of mental tests scores.* Reading, MA., Addison Wesley.
Martínez Arias, R. (1995). *Teoría de los tests psicológicos y educativos.* Madrid: Síntesis.
Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 49,* 529-544.
Muraki, E. y Bock, R.D. (1996). *Parscale 3.0.* Chicago: Scientific Software International.
Muñiz, J. (1990). *Teoría de la Respuesta a los ítems.* Madrid: Pirámide.
Olea, J. y Ponsoda, V. (1996). Tests adaptativos informatizados. En J. Muñiz (coord.): *Psicometría*. (pp. 729-783). Madrid: Universitas.
Owen, R.J. (1975). A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351-356.
Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics, 4,* 207-230.
Reise, S.P. y Henson, J.M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7,* 347-364.
Reise, S.P. y Henson, J.M. (2003). A discusion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81,* 93-104.
Reise, S.P. y Waller, N.G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14,* 45-58.
Robie, C., Zickar, M.J. y Schmit, M.J. (2001). Measurement equivalence between applicant and incumbent groups: an IRT analysis of personality scales. *Human Performance, 14,* 187-207**.**
Rojas, A.J. (2001). Pasado, presente y futuro de los tests adaptativos informatizados: entrevista con Isaac I. Bejar. *Psicothema, 13,* 685-690.
Rouse, S.V., Finger, M.S. y Butcher, J.N. (1999). Advances in clinical personality measurement: an item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment, 72,* 282-307.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scored. *Psychometrika Monograph, 17.*
Samejima, F. (1997). Graded response model. En W.J. Van der Linden y R.K. Hambleton (eds.): *Handbook of modern item response theory.* New York: Springer-Verlag.
Van der Linden, W.J. y Hambleton, R.K. (eds.) (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.
Zumbo, B., Pope, G.A., Watson, J.E. y Hubley, A.M. (1997). An empirical test of Roskam's conjecture about the interpretation of an ICC parameter in personality inventories. *Educational and Psychological Measurement, 57,* 963-969.