Apuntes de Psicología
2001, Vol. 19, número 1, pp. 11-28.
ISSN 0213-3334

Colegio Oficial de Psicólogos
(Andalucía Occidental) y
Universidad de Sevilla

Requerimientos, aplicaciones e investigación en tests adaptativos informatizados

Julio OLEA DÍAZ
Vicente PONSODA GIL
Javier REVUELTA MENÉNDEZ
Universidad Autónoma de Madrid
Pedro HONTANGAS BELTRÁN
Universidad de Valencia
Francisco J. ABAD GARCÍA
Universidad Autónoma de Madrid

Resumen

En este trabajo se resumen los principales requerimientos y aplicaciones de un Test Adaptativo Informatizado (TAI). Se comentan las principales diferencias entre los Tests Convencionales Informatizados y los TAIs, tanto en relación a los modelos psicométricos en que se sustentan como en las diferentes estrategias de presentación de ítems y estimación de los niveles de habilidad. Se describen además los principales resultados obtenidos por nuestro grupo de investigación, que ha desarrollado un conjunto de programas para la aplicación adaptativa de bancos de ítems, un TAI para evaluar el nivel de vocabulario inglés, y que además han realizado varios estudios para resolver algunos de los problemas técnicos y aplicados que presentan este tipo de pruebas.

Palabras clave: tests adaptativos informatizados, teoría de la respuesta al ítem.

Abstract

This article summarizes the main requirements and applications of Computerized Adaptive Testing (CAT). The work emphasizes the differences between CAT and the conventional computerized tests with regard to the psychometric models, the item selection strategies and the estimation of the examinee ability. The main results obtained by a Spanish research group in CATs are described. These results are: the development of software to implement adaptive testing, a CAT measuring English vocabulary in Spanish speaking subjects and several studies focusing in technical and applied issues of CAT.

Key words: computerized adaptive tests, item response theory.

Dirección del primer autor: Facultad de Psicología. Universidad Autónoma. Cantoblanco s/n. 28049 Madrid.
Correo electrónico: Julio.Olea@uam.es

El presente artículo pretende acercar al lector los fundamentos, utilidades, posibilidades y problemas que caracterizan a los Tests Adaptativos Informatizados (TAIs), un nuevo tipo de estrategia informatizada de evaluación que comienza a tener importantes aplicaciones (en determinados países) en contextos de evaluación psicológica y educativa.

Se presentará además una síntesis de los principales resultados de investigación obtenidos por un equipo de personas que, dentro del Grupo de Investigación en Medición Psicológica y Educativa -GIMPSE- (http://www.uv.es /~hontanga/), se han dedicado en los últimos años a la investigación sobre TAIs.

Lejos de la inevitable complacencia que provoca una síntesis de este tipo, lo que se pretende es ofrecer una primera aproximación a estos nuevos procedimientos de diseño y aplicación de tests psicométricos, así como las diversas estrategias de investigación empleadas para estudiar sus propiedades.

La idea fundamental de los TAIs es adaptar el test al nivel de cada evaluando. En los tests tradicionales se aplican todos los ítems a todos los evaluandos, lo que tiene el inconveniente de aplicar muchos items que informan poco sobre el nivel de algunas personas. Imaginemos, por ejemplo, un test de aptitud espacial. Aquellas personas de alta aptitud espacial acertarán sistemáticamente los ítems de dificultad baja y media, de modo que únicamente los ítems difíciles serán útiles para establecer diferencias entre este tipo de personas. Lo contrario sucedería con un grupo de personas de baja aptitud espacial. En un TAI únicamente se presentarán a cada persona los ítems que sean más informativos para estimar su nivel de capacidad.

Tests convencionales informatizados *vs.* tests adaptativos informatizados

Con los TAIs se pretende llegar a estimar el nivel de habilidad (conocimientos, aptitudes, destrezas, rasgos) de una persona de la manera más precisa y eficiente posible. Los objetivos no son muy distintos a los que se plantean con cualquier test psicométrico: medir bien los niveles de las personas y hacerlo con el menor número de ítems posible. Sí cambian bastante las maneras de conseguirlo, para lo cual dos herramientas se hacen imprescindibles: a) un modelo psicométrico que permita finalmente realizar estimaciones de los niveles de habilidad, y b) un ordenador que permita hacer con rapidez los complejos cálculos que es necesario realizar.

En relación a la primera de las herramientas, los TAIs se sustentan en modelos de la Teoría de la Respuesta al Ítem (TRI) que van a permitir la comparación de estimaciones de habilidad de evaluandos que han respondido a ítems diferentes, algo que representa sin duda un cambio fundamental de estrategias en la evaluación psicológica y educativa mediante tests.

El soporte informático resulta indispensable para realizar las estimaciones estadísticas y para tomar determinadas decisiones técnicas en el proceso de aplicación del test, relacionadas fundamentalmente con procedimientos específicos de selección de ítems.Antes de poner en funcionamiento un TAI, necesitamos: a) un banco de ítems calibrado, es decir, con propiedades psicométricas conocidas (al menos la dificultad de los ítems) y obtenidas desde un modelo de TRI; b) un algoritmo informático que controla la aplicación del test, compuesto por un criterio de arranque (que establece cómo comenzar la aplica-

ción o cómo seleccionar el primer ítem), una estrategia de continuación (que indica cómo seleccionar los siguientes ítems en función de las respuestas dadas a los ítems previos) y un criterio de parada (que indica cuando debe terminar el test).

Los TAIs más usuales comienzan con la presentación de un primer ítem de dificultad media. Después que el evaluando responde mediante el teclado o el ratón, se estima con un procedimiento estadístico un primer nivel de habilidad provisional, junto a un valor de precisión (muy baja en este punto del proceso, dado que la estimación de habilidad se ha realizado sólo con la respuesta a un ítem). Mediante el algoritmo de selección de ítems, se elige un segundo ítem apropiado para ese primer nivel estimado (por ejemplo uno de dificultad apropiada para ese nivel). El evaluando lo responde y, a partir de sus dos primeras respuestas, se estima un segundo nivel de habilidad provisional (ahora con mejor precisión) y se selecciona entre los ítems remanentes el que resulta más apropiado para ese segundo nivel estimado.

La TRI permite obtener en cada momento del proceso adaptativo la precisión con la que realizamos la estimación de habilidad, algo que iremos mejorando a medida que incrementemos el número de ítems presentados, y por tanto detener la aplicación cuando se alcance un nivel de precisión determinado, considerado aceptable y prefijado de antemano. También puede detenerse la aplicación cuando se presenta un número determinado de ítems, con lo que variará el nivel de precisión con el que medimos a los diferentes evaluandos.

Algunas empresas que se dedican al diseño y aplicación de tests comienzan a tener en sus catálogos Tests Convencionales Informatizados (TCI), normalmente para evaluar determinados rasgos de personalidad o ciertas aptitudes intelectuales.

Conviene, sin embargo, no confundir un TCI, que en definitiva es un test convencional presentado en soporte informático, con un TAI. Precisamente por el medio informático que emplean (que sirve para la presentación de ítems, la emisión de respuestas, la corrección del test y la elaboración de informes sobre el rendimiento) ambos tipos de tests informatizados comparten ya algunas ventajas: permiten estandarizar las condiciones de aplicación para todos los evaluandos, requieren menos tiempo para su aplicación, reducen la posibilidad de copia, pueden resultar más económicos en evaluaciones a gran escala, permiten obtener informes inmediatos sobre el rendimiento, minimizar los errores de corrección, presentar en la pantalla información dinámica, controlar los tiempos de respuesta y evaluar determinados procesos psicológicos difícilmente mensurables en un formato de papel y lápiz. Sin embargo, desde un punto de vista psicométrico, un TCI (y cualquier test convencional no informatizado) y un TAI tienen importantes diferencias, que se describen en la tabla 1.

Respecto al modelo psicométrico en que se fundamentan, los TCI utilizan los desarrollos de la Teoría Clásica de los Tests (TCT), mientras que los TAIs se sustentan necesariamente en alguno de los modelos de TRI.

Los TAIs más usuales se fundamentan en modelos unidimensionales para ítems dicotómicos, mientras que se están estudiando las posibilidades que tienen otros tests adaptativos sustentados en modelos de TRI politómicos (apropiados por ejemplo para tests de personalidad o escalas de

Tabla 1. Diferencias entre los Tests Convencionales Informatizados (TCI) y los Tests Adaptativos Informatizados (TAI).

	TCI	TAI
Modelo psicométrico	Teoría Clásica de los Tests	Teoría de la Respuesta al Ítem
Presentación de los ítems	Los mismos para todos los evaluandos	Diferentes. Los más apropiados para cada uno
Estimación del nivel de habilidad	Número de aciertos	Estimaciones máximo-verosímiles ó bayesianas
Precisión de las estimaciones	Coeficiente de fiabilidad Igual para todos los evaluandos	Error de medida para cada evaluando. Función de información.
Eficiencia del procedimiento	Mayor número de ítems Más tiempo de aplicación	Menor número de ítems Menos tiempo de aplicación
Motivación hacia el test según el nivel de los evaluandos	Frustración para niveles bajos, aburrimiento para niveles altos	Nivel similar de desafío para niveles bajos y altos
Sensación subjetiva de éxito	Mayor cuantos más aciertos se tienen	Baja para evaluandos con alto nivel

actitudes) y multidimensionales (ver Hontangas, Ponsoda, Olea y Abad, 2000). La mayoría de las diferencias entre ambos tipos de tests son consecuencias de esta distinta adscripción teórica.

Las dos propiedades fundamentales que aporta la TRI para los TAIs son la invarianza de los parámetros y el error típico de medida con que se estiman los niveles de habilidad. Sólo si el banco de ítems cumple los requisitos exigidos por la TRI (por ejemplo unidimensionalidad y ajuste de los ítems al modelo) se cumple en toda su extensión la propiedad de invarianza, según la cuál podemos estimar el nivel de habilidad presentando ítems diferentes. Sería difícil de imaginar cómo podemos comparar desde la TCT el rendimiento de dos personas a las que se presentan ítems que, por ejemplo, difieren en dificultad.

La utilización de modelos psicométricos diferentes tiene a su vez otro tipo de consecuencias. Por un lado, el nivel de habilidad de una persona en un test de rendimiento óptimo, según la TCT, se obtiene sumando el número de ítems acertados, y si acaso corrigiendo los posibles aciertos aleatorios; en la TRI, el nivel de habilidad se estima mediante métodos estadísticos (normalmente el de máxima verosimilitud o bayesianos), de tal forma que en algunos modelos la estimación del nivel de habilidad no sólo depende del número de aciertos obtenido, sino de las características de los ítems acertados y fallados (dificultad, discriminación, posibilidad de adivinación, etc.).

Por otro lado, la precisión desde la TCT se concibe como una medida global del test, similar para todos los niveles de

habilidad, y se obtiene mediante los oportunos procedimientos de cálculo del coeficiente de fiabilidad o del error típico de medida. En la TRI, la precisión es la varianza del estimador de habilidad (o su valor inverso, denominado como "Información") y se obtiene específicamente para cada estimación y, por tanto, para cada evaluando. Este aspecto es especialmente importante, ya que vamos a conocer cómo contribuye cada ítem y el test completo a la precisión con que estimamos los diferentes niveles de habilidad.

Algunas de las ventajas que tienen los procedimientos adaptativos tienen que ver con su eficiencia: se consiguen estimaciones de habilidad igual de precisas que en los TCI pero con un número de ítems aplicados sensiblemente inferior o, lo que en definitiva es lo mismo, pueden conseguirse estimaciones más precisas con el mismo número de ítems que un TCI. En contextos de evaluación donde es necesario aplicar muchos tests a grandes muestras (piénsese por ejemplo en situaciones de evaluación de tipo educativo o en procesos de selección de personal), la aplicación de TAIs puede representar un importante ahorro de tiempo y dedicación. Por otra parte, también mejoran la seguridad del test (y por tanto su validez) dado que una gran cantidad de ítems que se presentan a los evaluandos son diferentes.

Otras ventajas tienen que ver con el nivel motivacional con que los evaluandos afrontan un TAI: en un TCI de rendimiento óptimo, los evaluandos de alto nivel pueden experimentar cierto hastío cuando debe responder a items fáciles, mientras que los de bajo nivel pueden frustrarse al comprobar que muchos de los ítems no los saben; una cualidad de los TAIs es que presentan a cada evaluando los ítems más apropiados para su nivel. Ahora bien, esto puede resultar un inconveniente dado que en un TAI es normal acertar un poco más del 50 % de los ítems presentados, lo que rompe con la idea socialmente aceptada de que cuantos más aciertos tenemos mejor estamos haciendo un test y puede representar cierta sensación subjetiva de fracaso para determinadas personas con alto nivel de habilidad. Quizás un TAI no sea el mejor procedimiento de evaluación para personas con alto nivel de ansiedad.

No es éste el único inconveniente de los TAIs. Por ejemplo, la utilización de los modelos más usuales de TRI (los denominados modelos logísticos unidimensionales para ítems dicotómicos) exige la calibración del banco de ítems, esto es, la estimación de las propiedades psicométricas individuales de cada ítem, después de aplicarlo a muestras muy numerosas. Además, uno de los supuestos que se asume por estos modelos es el de unidimensionalidad del banco, es decir, que el rendimiento de las personas depende únicamente de una habilidad o rasgo subyacente, algo siempre difícil de conseguir cuando trabajamos con habilidades psicológicas. Otro tipo de inconvenientes, de carácter más técnico, serán comentados con posterioridad en este trabajo.

Afortunadamente tenemos ya importante documentación en castellano para profundizar en estos temas. Por ejemplo, quien desee una inmersión en la teoría y práctica de la TRI puede consultar los textos de López Pina (1995), Martínez Arias (1995), Muñiz (1996, 1997) o Santisteban (1990). Para ampliar los contenidos sobre TCIs pueden leerse los trabajos de Muñiz y Hambleton (1999) y Olea y Hontangas (1999). Barbero (1996, 1999) y Molina, Sanmartín y Pareja (1999) describen las

estrategias de diseño y uso de bancos de ítems. Sobre TAIs puede consultarse a Olea y Ponsoda (1996, 1998), Olea, Ponsoda, Revuelta, Hontangas y Suero (1999), Renom (1993) o Renom y Doval (1999). Hontangas (1999) proporciona información sobre el software disponible para la construcción de bancos de ítems, para la administración (adaptativa o no) de los tests y para los necesarios análisis psicométricos que es necesario realizar.

Algunas experiencias de investigación realizadas en España sobre tests informatizados y TAIs puede verse en el monográfico de la *Revista Electrónica de Investigación y Evaluación Educativa* (Olea y Ponsoda, 1999). Una reciente revisión sobre el estado actual de la investigación en TAIs puede consultarse en Hontangas *et al.* (2000).

Problemas a resolver y soluciones planteadas: una experiencia de investigación en TAIs

Elaboración de software y construcción de un banco de ítems

Cuando alguien quiere aplicar un banco de ítems de forma adaptativa puede recurrir a alguno de los programas comercializados (ver algunos en Hontangas, 1999) o puede elaborar los oportunos programas con un determinado lenguaje de programación. Los primeros trabajos del grupo (Revuelta, Ponsoda y Olea, 1993a; Ponsoda, Olea y Revuelta, 1994) consistieron en la elaboración de varios programas, denominados genéricamente como ADTEST, para la presentación adaptativa de ítems previamente calibrados mediante un modelo de TRI de 3 parámetros (este software puede obtenerse, sin cargo alguno, en la dirección http://www.uv.es/ ~hontanga/). El programa permite establecer el tiempo de exposición de los ítems en la pantalla y seleccionar un criterio de parada a priori (presentación de un número concreto de ítems o alcanzar una precisión determinada). Después de grabar el contenido de los diferentes ítems y sus correspondientes parámetros (*a*, discriminación; *b*, dificultad y *c*, pseudoazar), ADTEST permite la selección y aplicación de ítems del banco siguiendo el siguiente procedimiento:

a) Se inicia el proceso (de alguna manera hay que hacerlo) con una asignación aleatoria de habilidad para la persona, entre los valores centrales (de -1 a +1) de una distribución uniforme. Revuelta y Ponsoda (1997) proponen un procedimiento alternativo para la estimación del nivel de habilidad inicial.
b) Se obtiene una medida de la precisión que cada ítem del banco aporta para ese nivel de habilidad inicial.
c) Se elige el ítem más informativo (el más preciso) y se presenta como primer ítem al evaluando. Este ítem será uno con parámetro de dificultad próximo al nivel de habilidad asignado inicialmente.
d) Después de su respuesta, se estima un segundo nivel de habilidad provisional, según un procedimiento estadístico máximo-verosímil. Si el evaluando acierta el ítem, el nivel de habilidad estimado será superior al inicial; si lo falla, será inferior al que asignamos aleatoriamente.
e) Se presenta el ítem del banco más informativo para este segundo nivel de habilidad.
f) El proceso continúa hasta que se cumple el criterio de parada, es de-

cir, hasta que se presenta un número determinado de ítems o hasta que se alcanza determinado nivel de precisión prefijado de antemano. El programa graba la información fundamental de tipo psicométrico de todo el proceso de selección de ítems y estimación estadística.

En paralelo, se diseñó (Revuelta, Ponsoda y Olea, 1993b) un banco de 250 ítems de vocabulario inglés, cada uno de los cuales consistía en una palabra inglesa y cinco opciones de respuesta, entre las que se encontraba su correcta traducción al castellano. El banco se aplicó a una muestra heterogénea de nivel de inglés (estudiantes de secundaria, estudiantes universitarios y profesores universitarios), con un tiempo de exposición de cada ítem de 15 segundos, y se calibró según el modelo logístico de tres parámetros. Después del estudio psicométrico del banco se eliminaron 29 ítems, dado que no se ajustaban bien al modelo de TRI propuesto. Mediante técnicas factoriales se estudió el nivel de unidimensionalidad del banco, comprobando que la varianza explicada por el primer factor cumplía los requisitos mínimos exigibles. La función de información, que describe la precisión del banco para los diferentes niveles de habilidad, demostró su mayor eficacia para medir los niveles medios y medios-altos de vocabulario inglés.

El programa ADTEST fue modificado para que, además de la presentación adaptativa ya descrita, incluyera varios ítems de prueba (para que el evaluando se habituara al procedimiento de respuesta mediante el teclado) e información más concreta sobre el proceso de aplicación y los resultados. Diferentes modificaciones llevaron a lo que se denominó *APT-System*, que además de lo anterior permite utilizar un procedimiento bayesiano de estimación progresiva de la habilidad y tres estrategias diferentes de selección de ítems: una adaptativa (como la descrita), una segunda aleatoria (cada ítem se elige al azar del banco disponible) y otra autoadaptada, según la cual el evaluando puede elegir el nivel de dificultad antes de responder a cada uno de los ítems (ver más detalles en Olea y Ponsoda, 1996). En un apartado posterior se describirá el sentido de esta estrategia.

Propiedades psicométricas de las estimaciones

Ni el procedimiento informatizado, ni el modelo de TRI en que se sustenta, ni el algoritmo de presentación adaptativa representan en sí garantías del buen funcionamiento de un TAI. Como para cualquier otro tipo de test, o quizás más, un TAI debe someterse a una auditoría psicométrica lo más exahustiva posible para comprobar en qué grado refleja bien los niveles de rasgo de los evaluandos (precisión) y qué inferencias podemos realizar a partir de las estimaciones de habilidad que se obtienen (validez).

Precisión

Aunque algunos de los procedimientos más tradicionales para el estudio de la fiabilidad, por ejemplo el coeficiente de fiabilidad test-retest, pueden aplicarse al estudio de la fiabilidad de un TAI, otros resultan más novedosos y peculiares de los modelos de la TRI en que se sustentan. Básicamente, los estudios de precisión de un TAI pueden utilizar una metodología de simulación o llevarse a cabo con datos rea-

les. Los primeros consisten en simular las respuestas de muestras elevadas de evaluandos y tienen la ventaja de partir de parámetros (puntuaciones verdaderas) conocidos, con lo que podrá estudiarse la discrepancia entre las estimaciones que proporciona el TAI y los parámetros establecidos por el investigador. Los estudios empíricos de fiabilidad ponen a prueba el test en condiciones más realistas. Ambos tipos de estudios se realizaron para comprobar la precision de las estimaciones realizadas con ADTEST:

a) *Estudio de simulación.* Uno de los trabajos (Ponsoda, Olea y Revuelta, 1994) consistió en simular las respuestas de 3.750 evaluandos al TAI de vocabulario inglés. Para ello, se procedió de la siguiente forma:
   - Se establecieron 15 niveles de habilidad (parámetros) conocidos.
   - Se asignaron 250 evaluandos simulados a cada nivel.
   - Se tomaron como parámetros de los ítems los valores *a, b* y *c* obtenidos en el estudio de calibración del banco.
   - Conociendo la habilidad de un evaluando y los parámetros de un ítem, la TRI permite obtener la probabilidad que tiene el evaluando de acertar el ítem.
   - Para determinar si un evaluando simulado acierta o no un ítem, se selecciona un número aleatorio entre 0 y 1. Si la probabilidad de acierto que asigna el modelo es superior al número aleatorio se considera acertado, de lo contrario, se considera que el evaluando falla el ítem.
   - Mediante este procedimiento, se aplicó el TAI implementado en ADTEST con un criterio de parada mixto (aplicar 34 ítems o que la precisión alcanzara el valor 0.30). Esto significa que el TAI puede detenerse aplicando menos de 34 ítems si antes se alcanza la precisión deseada. Después de cumplir el criterio de parada, se obtuvo el nivel de habilidad estimado para cada evaluando, la precisión asociada a dicha estimación y el número de ítems aplicados.

   Como se conocen los parámetros de habilidad, resulta sencillo compararlos con las estimaciones finales que realiza el TAI. Por ejemplo, este trabajo permitió comprobar que el TAI estima bien los niveles medios y medio-altos, pero que sobrestima (asigna puntuaciones más elevadas) los niveles bajos. Se comprobó además que para la mayoría de los evaluandos se necesitó aplicar únicamente una media de 20 ítems para conseguir la precisión deseada.

b) *Estudio empírico.* En un segundo trabajo (Olea, Ponsoda, Revuelta y Belchí, 1996) se aplicó el TAI a una muestra de estudiantes de una academia de inglés. Cuando se conseguía el criterio de parada, y de manera imperceptible para los sujetos, se continuaba la presentación de los ítems remanentes del banco hasta que se agotaban los 221. La correlación entre las estimaciones de habilidad realizadas con el TAI y las obtenidas a partir de las respuestas al banco completo fue 0.9. Significa esto que el orden que asignamos con

el TAI a los sujetos en nivel de vocabulario inglés es muy parecido al que asignamos a partir de sus respuestas al banco completo.

Validez

Seguramente no hay muchas dudas que un banco de ítems como el descrito mide nivel de vocabulario inglés, pero para demostrarlo se plantearon dos hipótesis diferentes:

a) El TAI debería ser capaz de discriminar el nivel de inglés de diversos niveles académicos.
b) Las puntuaciones estimadas deberían correlacionar con otras pruebas de inglés que miden otro tipo de destrezas relacionadas con el dominio de este idioma.

Para contrastar la primera hipótesis, se realizó un ANOVA donde la variable independiente era el nivel educativo (1°, 2° y 3° de BUP, COU, estudiantes universitarios y profesores universitarios) y la variable dependiente los niveles de vocabulario estimados; prácticamente todas las comparaciones de medias entre diferentes niveles educativos resultaron significativas, lo que quiere decir que el test proporciona puntuaciones medias más elevadas cuanto mayor es el nivel educativo de los evaluandos.

Para contrastar la segunda hipótesis, se aplicó a un grupo de estudiantes el TAI junto al *Oxford Placement Test*, una prueba que proporciona medidas de nivel de gramática inglesa y *listening*; se obtuvieron relaciones lineales significativas con ambas medidas, mayores para las puntuaciones de gramática.

Efectos en los evaluandos: tests autoadaptados informatizados y revisión de respuestas

Uno de los problemas ya indicados que pueden generar los TAIs es la tasa de aciertos que permite (algo superior al 50 %), lo que puede no resultar óptimo desde un punto de vista motivacional para determinados evaluandos (por ejemplo los más ansiosos). Para intentar paliar algunos de estos efectos negativos, que por otra parte podría incidir en su rendimiento, se propusieron a finales de los 80 los denominados Tests Autoadaptados Informatizados -TADIs- (Wise, 1999; Wise, Ponsoda y Olea, en prensa). Este tipo de tests se asemejan bastante a los TAIs pero tienen una diferencia importante: el evaluando puede elegir el nivel de dificultad en que se quiere ubicar antes de responder a un ítem. Para que esto sea posible, se divide el banco en k (normalmente entre 5 y 9) niveles o estratos de dificultad. Antes de responder un ítem, el evaluando se sitúa en un nivel y el algoritmo elige el ítem más apropiado de los que se incluyen en dicho nivel. Después que un individuo emite su respuesta, es importante proporcionarle información sobre el resultado (acierto o error), para que así pueda ubicarse en el siguiente ítem en la categoría que desee. ¿Qué se pretende con esto? La idea es que el evaluando participe de esta forma en la sesión de evaluación para que ajuste su nivel de ansiedad y el confort general ante la prueba al punto óptimo que facilite su rendimiento, todo ello, idealmente, manteniendo las propiedades psicométricas de las estimaciones que se hacen con un TAI. En algunos trabajos se ha encontrado que los TADIs reducen la ansiedad de los sujetos ante la evaluación, que atenúan o anulan

las relaciones negativas entre ansiedad y rendimiento, y que permiten rendir mejor a ciertas personas.

El equipo realizó 4 estudios diferentes sobre TADIs, algunos junto al profesor Wise de la *James Madison University*, dividiendo el banco de inglés en 5 ó 7 categorías ordenadas de dificultad:

Estudio 1 (Ponsoda, Wise, Olea y Revuelta, 1997).

Se compararon los efectos de 4 estrategias diferentes de selección de ítems: adaptativa, autoadaptada, aleatoria y combinada (los primeros ítems de forma adaptativa, los últimos de forma autoadaptada). Antés y después de cada test, se aplicó una versión informatizada del cuestionario de ansiedad-estado de Spielberger. No se encontraron diferencias significativas en ansiedad entre las 4 estrategias, aunque el tamaño del efecto entre el TAI y el TADI fue comparable al de investigaciones previas. Se obtuvo mejor precisión en el TAI y mayores tasas de acierto en el TADI.

Estudio 2 (Olea, Ponsoda, Felipe y Carretié, 1998).

En el trabajo precedente no se obtuvieron niveles de ansiedad pre-test elevados en los sujetos, por lo que se planificó un segundo trabajo con estudiantes de alta ansiedad de evaluación, donde se midiera (además de la ansiedad estado con el cuestionario) la actividad electrodérmica de los sujetos durante la sesión de evaluación. No se obtuvieron diferentes tasas de aciertos entre las estrategias adaptativa y autoadaptada. Tampoco se obtuvieron diferencias significativas entre el TAI y el TADI en las dos medidas de ansiedad. Los resultados de estos dos primeros trabajos no apoyan la hipótesis del descenso de la ansiedad en los TADIs, pero algún indicio se encontró de que tal descenso podría darse cuando en el TADI se consiguen tasas de acierto elevadas (por ejemplo en evaluandos que eligen categorías asequibles de dificultad, lo que les permite tener más aciertos).

Estudio 3 (Ponsoda, Olea, Rodríguez y Revuelta, 1999).

Ante la evidencia de los dos estudios previos sobre la posible incidencia de la tasa de acierto (y no tanto de permitir elegir la dificultad) en los niveles de ansiedad, se planificó un estudio experimental donde se manipuló la dificultad tanto en el TAI como en el TADI. Se establecieron cuatro grupos experimentales, a los que se aplicó respectivamente uno de estos 4 tests: TAI-fácil, TAI-difícil, TADI-fácil, TADI-difícil. En las dos condiciones fáciles se obtuvieron mayores tasas de acierto y menor ansiedad postest, lo que revela cierta incidencia de la *sensación subjetiva de éxito* en el descenso de los niveles de ansiedad ante la evaluación.

Estudio 4 (Hontangas, Olea y Ponsoda, 1999; Hontangas, Ponsoda, Olea y Wise, 2000).

El descenso de ansiedad que se ha obtenido en algunos estudios sobre TADIs podría depender de las diferentes estrategias de elección de la dificultad que tienen los evaluandos en este tipo de tests. En estos trabajos, cada evaluando fue clasificado en una de las siguientes estrategias: flexible (cuando acierta sube de nivel, cuando falla baja), tolerante al fracaso (se queda en el mismo nivel si falla, sube de

nivel si acierta), intolerante al fracaso (se queda en el mismo nivel si acierta, baja de nivel si falla) e inflexible (se mantiene en el mismo nivel, independientemente de que acierte o falle). Los evaluandos de bajo nivel de ansiedad tienden a seguir una estrategia tolerante al fracaso, mientras que los de alta ansiedad siguen en mayor grado estrategias flexibles o inflexibles. Sólo los que siguieron estas dos últimas estrategias obtuvieron un descenso significativo de la ansiedad después de responder al TADI.

Como puede comprobarse, la investigación sobre tests informatizados no se centra exclusivamente en conseguir buenas estimaciones, sino en mejorar las condiciones en que los evaluandos deben responderlos. En algunos trabajos se ha obtenido un mayor nivel de rendimiento en los TADIs que en los TAIs, lo que plantea un importante interrogante, no tanto de índole psicométrico como ético, sobre la interferencia del nivel de ansiedad en este tipo de pruebas (y en otras....los alumnos de COU y bachillerato acaban de hacer el examen de selectividad): ¿las personas más ansiosas, que rinden por debajo de su nivel en situaciones reales de evaluación donde se juegan mucho, deberían ser evaluadas con otro tipo de procedimientos que mitigasen estos efectos perniciosos en el rendimiento?

Otra cuestión relacionada con el confort de los sujetos en los TAIs es que no permiten que los evaluandos revisen y modifiquen sus respuestas dadas a los ítems, lo cual se percibe como algo incómodo e injusto. La razón fundamental de no permitir la revisión de respuestas en los TAIs es que podrían darse lo que se denomina como estrategias ilegítimas de respuesta, como por ejemplo fallar deliberadamente los ítems en la primera aplicación para resolverlos correctamente después de la revisión. En un reciente trabajo empírico (Olea, Revuelta, Ximénez y Abad, 2000) se establecieron dos tipos de condiciones: a) revisión no permitida, y b) revisión permitida al final del TAI, según la cual los evaluandos podían cambiar las respuestas que inicialmente dieron. Entre otras cosas se comprobó que los evaluandos de la condición b) obtuvieron un descenso medio en ansiedad después de realizar la prueba, mientras que los de la condición a) incrementaron su nivel de ansiedad. Después de la revisión se incrementó significativamente el nivel de habilidad estimada y el número medio de aciertos.

El problema del control de la exposición de los ítems

Ya hemos indicado que, según el procedimiento de máxima información que guía el proceso de selección de ítems en un TAI, después de una respuesta se elige el ítem más informativo para el nivel de habilidad provisional estimado. Normalmente, esto lleva a presentar los ítems más discriminativos del banco (los que tienen mayor parámetro *a*) mientras que los menos discriminativos no aparecen en casi ninguno de los tests que se aplican a evaluandos con diferente nivel de habilidad. En definitiva, esto sería algo así como trabajar con un banco de dimensiones más reducidas del que auténticamente se dispone. El control de la tasa de exposición tiene que ver con establecer procedimientos para conseguir un objetivo fundamental: reducir la exposición de los items más discriminativos, pues no hacerlo puede representar una amenaza para la validez del test y para la seguridad del banco (por ejemplo si se difunden entre los evaluandos en procesos de

evaluación a gran escala). Esto tiene como consecuencia inmediata un incremento de la exposición de los ítems menos discriminativos, lo que puede suponer entre otras cosas una garantía de la validez de contenido del test y rentabilizar la inversión realizada con la construcción del banco. Se trata de dar con el procedimiento más eficaz, en el sentido de conseguir ambos objetivos sin que el TAI pierda mucho en precisión.

Varios trabajos (Revuelta y Ponsoda, 1996; Revuelta y Ponsoda, 1998, Revuelta, Ponsoda y Olea, 1999) se han orientado a estudiar mediante simulación la eficacia de distintos procedimientos, unos ya descritos por otros autores y otros propuestos originalmente. Entre los propuestos por otros autores se ha estudiado, por ejemplo, el efecto que tiene: a) presentar al azar un item entre los 5 más informativos, b) ir reduciendo el número de ítems más informativos entre los que escoger (primero entre 5, luego entre 4, 3, 2 y 1), c) fijar un porcentaje máximo de exposición en los diferentes tests, y d) incluir un parámetro de control, establecido para cada ítem mediante métodos de simulación.

En el primero de estos trabajos se propuso un nuevo método, denominado como *progresivo*, según el cual los primeros ítems se seleccionan de forma casi aleatoria, y cada vez se concede más importancia a la información proporcionada por cada ítem. En el segundo se plantean dos estudios de simulación en los que se establecen diferentes tipos de condiciones (longitud fija y variable del TAI, diferentes distribuciones para los parámetros de los ítems) y se analiza la eficacia del método progresivo respecto a los otros y al procedimiento usual de un TAI sin control de la tasa de exposición. Sin que produzca un decremento considerable de la precisión del TAI, el método progresivo parece funcionar bien para evitar la infrautilización de ítems.

Generación automática de ítems

En los últimos años se está produciendo un importante acercamiento entre la psicología cognitiva y la psicometría (ver Olea, Ponsoda y Prieto, 1999; Prieto y Delgado, 1999; Revuelta y Ponsoda, 1999), en el sentido de estudiar desde modelos matemáticos los procesos cognitivos que intervienen en la resolución de los ítems de un test. Un claro ejemplo de esta aproximación lo representan los denominados *modelos componenciales*, que sirven para estimar la dificultad de los ítems a partir de los procesos implicados en su resolución. Imaginemos, por ejemplo, los ítems de un test de rotación de figuras tridimensionales en el espacio, que plantean un estímulo determinado y varias opciones de respuesta de las que sólo una es correcta (el estímulo rotado ciertos grados). Si la opción correcta resulta de girar sólo 20° el estímulo, el ítem será más fácil que otro donde llegar a la solución correcta supone un giro de 220° y además dar la vuelta al estímulo (porque representa una imagen en espejo). Además, es seguro que la complejidad del estímulo tiene también relación con su dificultad. Pues bien, si se delimitan de forma adecuada desde un modelo cognitivo concreto los procesos que intervienen en las tareas planteadas, y si sabemos el número de veces que deben repetirse los procesos para llegar a la solución correcta, podría programarse el ordenador para que genere automáticamente ítems de dificultad conocida.

¿Qué ventajas tendría la Generación Automática de Ítems (GAI)? En primer lugar, que no es poco, evitaría los costosos

procesos de aplicación de bancos de ítems a grandes muestras para conocer sus propiedades psicométricas. En segundo lugar, incrementaría la validez diagnóstica de un test, dado que podríamos detectar los déficits cognitivos que son responsables de los fallos. En tercer lugar, se aportan pruebas sobre la validez de constructo del test. Finalmente, podrían diseñarse TAIs a partir de la dificultad predicha por el modelo.

Las primeras experiencias del grupo de investigación sobre GAI se recogen en dos trabajos fundamentales (Revuelta y Ponsoda, 1998; Real, Olea, Ponsoda, Revuelta y Abad, 1999). El primero de ellos estudia los procesos intervinientes en un test comercializado de análisis lógico, el DA5. El segundo analiza los procesos que se requieren para resolver un test de matemáticas para alumnos de secundaria. Ambos coinciden en el siguiente esquema de trabajo:

a) Delimitación de los componentes que intervienen en la dificultad de los ítems.
b) Descripción de las operaciones mentales que deben ejecutarse (cantidad y tipo) para resolver los ítems.
c) Estudio psicométrico del test: estimación de la dificultad de los ítems mediante un modelo de TRI y mediante un modelo componencial. Este segundo permite estimar el peso que tienen los diferentes componentes o procesos delimitados en la correcta resolución de los ítems.
d) Comprobar la correspondencia entre los parámetros de dificultad estimados mediante ambos modelos.

Después de comprobar la adecuación del modelo componencial, puede programarse el sistema para generar ítems que contengan determinadas combinaciones de los componentes y que, por tanto, tengan diferente dificultad. Por ejemplo, en el primero de los trabajos se generaron un total de 4.232 ítems diferentes, que puede considerarse la *población* de ítems que sirva de banco para la aplicación adaptativa del test de análisis lógico.

Como el lector puede aventurar, este campo de trabajo no está exento de problemas importantes que deberán resolverse de forma progresiva: muchas veces no disponemos de modelos cognitivos apropiados para analizar los componentes que intervienen en la resolución de un test, otras veces son los modelos psicométricos componenciales los que no se encuentran desarrollados en toda su extensión o no consideran efectos que son importantes, como es el aprendizaje o las diferencias individuales en procesamiento. Sin embargo, las potencialidades de estas estrategias, que pretenden aunar los avances en psicología, psicometría e informática, son en este momento más que interesantes.

Evaluación final y prospectiva

Muy probablemente alguien piense, con parte de razón, que todas estas cosas no son sino juegos florales o problemas que intentan resolver conciencudos investigadores universitarios alejados de las necesidades reales de evaluación que se plantean en contextos clínicos, organizacionales o educativos. Esas personas deberían saber, al menos, que en Estados Unidos y algunos paises europeos se aplican TAIs en contextos de evaluación psicológica y educativa a gran escala para evaluar diverso tipo de conocimientos, aptitudes o destrezas. En diferentes contextos de evaluación educativa se aplican de forma

adaptativa tests como el *Graduate Record Examinations* (GRE), el *Graduate Management Admissions Tests* (GMAT), el *Differential Aptitude Tests* (DAT), el *COM PASS Placement Tests* o el *Test of English as a Foreign Language* (TOEFL). Diversos colegios profesionales han impulsado la investigación y aplicación de TAIs en exámenes de licenciatura y certificación (por ejemplo para enfermería, medicina). En contextos de recursos humanos, fundamentalmente relacionados con los procesos de selección de personal que se realizan en fuerzas armadas profesionales, se utilizan versiones adaptativas de tests muy conocidos, como el *Armed Services Vocational Aptitude Battery* (ASVAB) en USA o algunos de los subtests de la batería *Micropat* en Inglaterra. Comienzan a emplearse TAIs de propósito tan diverso como evaluar destrezas musicales, el nivel de inglés, los conocimientos requeridos para admitir a los estudiantes en determinadas facultades de derecho, las actitudes hacia el consumo de alcohol o el nivel de calidad de vida relacionado con la salud. Eso sí, siempre en otros países.

En nuestro país nos conformamos por ahora con disponer de algunas versiones informatizadas de tests psicométricos convencionales de aptitudes o personalidad. Resulta evidente que existe una importante distancia entre la investigación psicométrica que se realiza en las Facultades de Psicología o de Educación y la propuesta de instrumentos informatizados de evaluación que tengan utilidad de diverso tipo para los profesionales. ¿Por qué? Seguramente las razones son muchas, pero sólo avanzamos algunas:

a) Los profesionales de la psicometría no saben detectar en toda su extensión las necesidades de tipo aplicado que tienen los psicólogos y educadores en su trabajo cotidiano.
b) Los investigadores no saben tampoco transmitir las ventajas económicas que pueden representar (al menos a medio plazo) los tests informatizados en general, y los TAIs en particular, en determinadas aplicaciones a gran escala de tests.
c) La complejidad matemática de las estimaciones que se realizan mediante la TRI no facilita su entendimiento y uso por profesionales no expertos en psicometría.
d) Queda por demostrar que los tests informatizados de procesos cognitivos resultan mejores predictores de rendimientos que los tests fundamentados en teorías del rasgo.
e) Algunas personas e instituciones tienen todavía evidentes reticencias al uso de los ordenadores como medio de administración de tests.
f) No existe un valor social positivo hacia el uso de los tests como instrumentos de evaluación del conocimiento (piénsese cómo se realizan exámenes tan importantes como la prueba de acceso a la universidad, el de Médicos Internos Residentes u otros).
g) Quizás como consecuencia de esto último, no disponemos de empresas (como el *Educational Testing Service* o el *American College Testing*) que se encarguen del diseño de tests específicos para objetivos muy concretos.

Nos consta que se están produciendo importantes avances para cambiar esta situación. El Colegio Oficial de Psicólogos está trabajando en el establecimiento de

una guía para la evaluación de tests comercializados, en la que se pueden incluir criterios para valorar tests informatizados (Prieto y Muñiz, 2000). En diversas facultades se imparte docencia sobre TRI y sobre tests informatizados, se investiga sobre estos temas y se publican textos divulgativos. En algunas se han desarrollado ya, por ejemplo, tests informatizados de aptitudes espaciales para su aplicación en el ejército del aire (Gerardo Prieto y colaboradores de la Universidad de Salamanca), varios programas para la aplicación adaptativa de tests (Jordi Renom y colaboradores de la Universidad de Baracelona) o TAIs para evaluar conocimientos y destrezas relacionados con materias de la ESO (Eduardo García y colaboradores de la Universidad de Sevilla). Por nuestra parte, disponemos ya de un banco de 635 ítems de gramática inglesa, confeccionado por especialistas en filología, para desarrollar un sistema adaptativo que resulte válido para estimar las diferentes destrezas que tienen que ver con el dominio del idioma inglés.

Referencias

Barbero, M.I. (1996). Bancos de ítems. En J. Muñiz (Coord.). *Psicometría*. Madrid: Universitas.

Barbero, M. I. (1999). Gestión informatizada de bancos de ítems. En J. Olea, V. Ponsoda y G. Prieto (Eds.). *Tests informatizados: Fundamentos y aplica ciones*. Madrid: Pirámide.

Hontangas, P. (1999). Software para la construcción y administración de tests informatizados. En J. Olea, V. Ponsoda y G. Prieto (Eds.). *Tests informatizados: Fundamentos y aplicaciones*. Madrid: Pirámide.

Hontangas, P., Olea, J. y Ponsoda, V. (1999). Elección de la dificultad en los tests autoadaptados informatizados: Un estudio piloto. *Revista Electrónica de Investigación y Evaluación Educativa, 4 (2)* [http://www2.uca.es/ RELIEVE /V4N2].

Hontangas, P., Ponsoda, V., Olea, J. y Abad, F.J.(2000). Los tests adaptativos informatizados en la frontera del siglo XXI: una revisión. *Metodología de las Ciencias del Comportamiento, 2 (2),* 183-216.

Hontangas, P., Ponsoda, V., Olea, J. y Wise, S.L. (2000). The choice of item difficulty in self-adapted testing. *European Journal of Psychological Assess ment, 16 (1),* 3-12.

López Pina, J.A. (1995). *Teoría de Respuesta a los Ítems: Fundamentos*. Murcia: DM-PPU.

Martínez Arias, M.R. (1995). *Psicometría: Teoría de los tests psicológicos y educativos*. Madrid: Síntesis.

Molina, J.G., Sanmartín, J. y Pareja, I. (1999). Desarrollo de un sistema informático orientado a la construcción y gestión de bancos de ítems. En J. Olea, V. Ponsoda y G. Prieto (Eds.). *Tests informati zados: Fundamentos y aplicaciones*. Madrid: Pirámi de.

Muñiz, J. (Coord.) (1996). *Psicometría*. Madrid: Universitas.

Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems*. Madrid: Pirámide.

Muñiz, J. y Hambleton, R.K. (1999). Evaluación psicométrica de los tests informatizados. En J. Olea, V. Ponsoda y G. Prieto (Eds.). *Tests informatizados: Fundamentos y aplicaciones*. Madrid: Pirámide.

Olea, J. y Hontangas, P. (1999). Tests informatizados de primera generación. En J. Olea, V. Ponsoda y G. Prieto (Eds.). *Tests informatizados: Fundamentos y aplicaciones.* Madrid: Pirámide.

Olea, J. y Ponsoda, V. (1996). Tests adaptativos informatizados. En J. Muñiz (Coor.). *Psicometría.* Madrid: Universitas.

Olea, J. y Ponsoda, V. (1998). Evaluación informatizada en contextos de aprendizaje. En C. Vizcarro y J.A. León (Eds.). *Nuevas tecnología para el aprendizaje.* Madrid: Pirámide.

Olea, J. y Ponsoda, V. (1999). Tests informatizados y adaptativos informatizados: investigación en España. *Revista electrónica de investigación y evaluación educativa, 4 (2),* [http:/ /www2.uca.es/ relieve/v4n2_pre.htm].

Olea, J., Ponsoda, V., Felipe, L. y Carretié, L. (1998). Estrategias de evaluación informatizada y ansiedad: Un estudio comparativo en una muestra de alto riesgo. *Ansiedad y Estrés,* 4, 1, 71-79.

Olea, J., Ponsoda, V. y Prieto, G. (Eds.) (1999). *Tests informatizados: Fundamentos y aplicaciones.* Madrid: Pirámide.

Olea, J., Ponsoda, V., Revuelta, J. y Belchí, J. (1996). Propiedades psicométricas de un test adaptativo informatizado de vocabulario inglés. *Psicológica, 55,* 61-73.

Olea, J., Ponsoda, V., Revuelta, J., Hontangas, P. y Suero, M. (1999). Investigación en tests adaptativos informatizados. En J. Olea, V. Ponsoda y G. Prieto (Eds.). *Tests informatizados: Fundamentos y aplicaciones.* Madrid: Pirámide.

Olea, J., Revuelta, J., Ximénez, C. y Abad, F.J. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive tests. *Psicológica, 21,* 157-173.

Ponsoda, V., Olea, J. y Revuelta, J. (1994). ADTEST: A computer-adaptive test based on the maximun information principle. *Educational and Psycholo gical Measurement, 54 (3),* 680-686.

Ponsoda, V., Olea, J., Rodríguez, M.S. y Revuelta, J. (1999). The effects of test difficulty manipulation in computerized adaptive testing and self-adapted testing. *Applied Measurement in Education, 12,* 167- 184.

Ponsoda, V., Wise, S.L., Olea, J. y Revuelta, J. (1997). An investigation of self-adapted testing in a spanish high school population. *Educational and Psychological Measurement, 57 (2),* 210-221.

Prieto, G. y Delgado, A.R. (1999). Medición cognitiva de las aptitudes. En J. Olea, V. Ponsoda y G. Prieto (Eds.). *Tests informatizados: Fundamentos y aplicaciones.* Madrid: Pirámide.

Prieto, G. y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 77,* 65-75.

Real, E., Olea, J., Ponsoda, V., Revuelta, J. y Abad, F.J. (1999). Análisis de la dificultad de un test de matemáticas mediante un modelo componencial. *Psicológica, 20,* 121-134.

Renom, J. (1993). *Tests adaptativos computerizados: Fundamentos y aplicaciones.* Barcelona: PPU.

Renom, J. y Doval, E. (1999). Tests adaptativos informatizados: Estructura y desarrollo. En J. Olea, V. Ponsoda y G. Prieto (Eds.). *Tests informatizados: Fundamentos y aplicaciones.* Madrid: Pirámide.

Revuelta, J. y Ponsoda, V. (1996). Métodos sencillos para el control de la tasa de exposición en tests adaptativos informatizados. *Psicológica, 17,* 161-172.

Revuelta, J. y Ponsoda, V. (1997). Una solución a la estimación inicial en los tests adaptativos informatizados. *Revista Electrónica de Metodología Aplicada, 2 (2),* 1-6.

Revuelta, J. y Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35,* 311-327.

Revuelta, J. y Ponsoda, V. (1999). Generación automática de ítems. En J. Olea, V. Ponsoda y G. Prieto (Eds.). *Tests informatizados: Fundamentos y aplicaciones.* Madrid: Pirámide.

Revuelta, J., Ponsoda, V. y Olea, J. (1993a). ADTEST: A maximun information Computerized Adaptive Test. *Applied Psychological Measurement, 17,* 1-28.

Revuelta, J., Ponsoda, V. y Olea, J. (1993b). Un test adaptativo informatizado de vocabulario inglés: Descripción del programa. *Psicológica, 14,* 347-354.

Revuelta, J., Ponsoda, V. y Olea, J. (1999). Métodos para el control de las tasas de exposición en tests adaptativos informatizados. *Revista electrónica de investigación y evaluación educativa.* [http://www2.uca.es/RELIEVE/V4N2].

Santisteban, C. (1990). *Psicometría: Teoría y práctica en la construcción de tests.* Madrid: Norma.

Wise, S.L. (1999). Tests autoadaptados informatizados: Fundamentos, resultados de investigación e implicaciones para la aplicación práctica. En J. Olea, V. Ponsoda y G. Prieto (Eds.). *Tests informatizados: Fundamentos y aplicaciones.* Madrid: Pirámide.

Wise, S.L., Ponsoda, V. y Olea, J. (en prensa). Self-Adapted Testig: An overview. *International Journal of Continuing Engineering Education.*