Papeles del Psicólogo, 2010. Vol. 31(1), pp. 94-107
http://www.cop.es/papeles

TESTS INFORMATIZADOS Y OTROS NUEVOS TIPOS DE TESTS

Julio Olea[1], Francisco J. Abad[1] y Juan R. Barrada[2]
[1]*Universidad Autónoma de Madrid.* [2]*Universidad Autónoma de Barcelona*

Recientemente se ha producido un considerable desarrollo de los tests adaptativos informatizados, en los que el test se adapta progresivamente al rendimiento del evaluando, y de otros tipos de tests: a) los test basados en modelos (se dispone de un modelo o teoría de cómo se responde a cada ítem, lo que permite predecir su dificultad), b) los tests ipsativos (el evaluado ha de elegir entre opciones que tienen parecida deseabilidad social, por lo que pueden resultar eficaces para controlar algunos sesgos de respuestas), c) los tests conductuales (miden rasgos que ordinariamente se han venido midiendo con autoinformes, mediante tareas que requieren respuestas no verbales) y d) los tests situacionales (en los que se presenta al evaluado una situación de conflicto laboral, por ejemplo, con varias posibles soluciones, y ha de elegir la que le parece la mejor descripción de lo que el haría en esa situación). El artículo comenta las características, ventajas e inconvenientes de todos ellos y muestra algunos ejemplos de tests concretos.
Palabras clave: *Test adaptativo informatizado, Test situacional, Test comportamental, Test ipsativo y generación automática de ítems.*

The paper provides a short description of some test types that are earning considerable interest in both research and applied areas. The main feature of a computerized adaptive test is that in despite of the examinees receiving different sets of items, their test scores are in the same metric and can be directly compared. Four other test types are considered: a) model-based tests (a model or theory is available to explain the item response process and this makes the prediction of item difficulties possible), b) ipsative tests (the examinee has to select one among two or more options with similar social desirability; so, these tests can help to control faking or other examinee's response biases), c) behavioral tests (personality traits are measured from non-verbal responses rather than from self-reports), and d) situational tests (the examinee faces a conflictive situation and has to select the option that best describes what he or she will do). The paper evaluates these types of tests, comments on their pros and cons and provides some specific examples.
Key words: *Computerized adaptive test, Situational test, Behavioral test, Ipsative test and y automatic item generation.*

Hace un par de años que varios historiadores de la Psicología Española (Quintana, Bitaubé y López-Martín, 2008) rescataron y editaron unos "Apuntes para un curso de Psicología aplicada a la selección profesional", elaborados en 1924 por el doctor Rodrigo Lavín como material docente de su cátedra de Psicología Experimental. Esta auténtica joya representa una de las primeras veces que en España se habla sistemáticamente de los tipos y usos de los tests. Decía ya entonces el autor: "*Como la observación nos da muy pocos datos utilizables y la conversación o entrevista no basta para descubrir las habilidades de los solicitantes, es necesario recurrir a los tests. Se puede decir que estamos en el comienzo de los tests y, a pesar de eso, hay un desarrollo extraordinario de ellos; ello indica lo que sucederá andando el tiempo*". Hablaba el autor de que existían entonces tests de capacidades o habilidades, tanto generales como específicas, y que en la selección profesional eran de especial importancia los tests de fuerza, de resistencia a la fatiga, de control motor y de capacidades mentales (atención, sensación y percepción, imaginación e inteligencia general).

Andado el tiempo hasta hoy, el desarrollo de los tests ha sido extraordinario, como anticipaba Lavín, tanto en la variedad como en la complejidad. Prueba de ello es que las clasificaciones simples de los tipos de tests (por ejemplo, la que distinguía entre "tests impresos" y "tests manipulativos" o las que se referían al diferente contenido de las pruebas) se han quedado obsoletas por la presencia de nuevos tipos de tests que eran difíciles de prever en el pasado. Todo ello se ha debido a distintos factores:

- ✔ *Avances técnicos*. El desarrollo de los modelos psicométricos que sustentan las propiedades métricas de los tests y la evolución y abaratamiento de la tecnología informática nos ha permitido incorporar nuevos atributos psicológicos al catálogo de lo medible; también ha permitido incrementar la eficiencia de las aplicaciones e incluir nuevas funcionalidades, como son la generación automática de ítems, la aplicación adaptativa de un test o la corrección automática de respuestas complejas.

Correspondencia: Julio Olea. *Facultad de Psicología. Universidad Autónoma de Madrid. Calle Iván Pavlolv 6. 28049 Madrid. España. E-mail: Julio.olea@uam.es*

- ✔ *Nuevas demandas sociales.* En España, aunque todavía con cierta lejanía respecto a otros países, tanto los profesionales de la Psicología como otros responsables de organizaciones públicas y privadas confían cada vez más en la utilidad de los tests para conseguir ciertos objetivos aplicados, como lo prueba el artículo de Muñiz y Fernández-Hermida (2010) de este mismo número. Pero no sólo se incrementa el uso de los tests "clásicos" como el WAIS o el 16PF. En una sociedad cada vez más sensible a la evaluación de los resultados de las intervenciones y a la acreditación de competencias individuales e institucionales, se ha ampliado mucho el tipo de atributos psicológicos que se precisa medir. Mientras que hace unos años las aplicaciones fundamentales se ceñían a tests de capacidades cognitivas o pruebas de personalidad, cada vez son más los profesionales que exigen buenos tests para objetivos específicos.
- ✔ *Mayor exigencia de calidad.* Cada vez son más importantes las consecuencias que para las personas y las organizaciones tienen las puntuaciones en los tests. Por ello, también es mayor la exigencia psicométrica a la que sometemos a las puntuaciones de los tests. El ineludible requisito de "medir bien" y la necesidad de afrontar problemas singulares en ciertos contextos de evaluación (como puede ser el falseamiento de las respuestas en contextos de selección) está impulsando el desarrollo de nuevos tipos de tests y nuevos modelos psicométricos para estudiar las garantías que ofrecen sus aplicaciones.

TESTS INFORMATIZADOS

Se van incrementando progresivamente los tests cuyos ítems se presentan, se responden y puntúan en un ordenador, lo que ha representado cambios y avances importantes en contextos aplicados de evaluación psicológica y educativa. Para Davey (2005): *"En las últimas dos décadas los tests informatizados han pasado de ser un procedimiento experimental a ser empleado por cientos de programas de evaluación que evalúan a millones de personas cada año"* ... *"ser evaluado mediante un ordenador puede pronto llegar a ser incluso más natural que ser evaluado en papel"* (p. 358).

Estrictamente hablando, un test informatizado debe cumplir dos requisitos (Olea, Ponsoda y Prieto, 1999): a) que se conozcan las propiedades psicométricas de los ítems que lo integran, estimadas a partir de un modelo matemático, y b) que los ítems se presenten y respondan en un ordenador. El primero de estos requisitos excluye de la consideración como "test informatizado" a muchos de los tests que sin las oportunas garantías se ofrecen en Internet.

El ordenador permite aplicar los tests de diversos modos. Existen en primer lugar los *tests fijos informatizados.* En estos tests los ítems se aplican en la misma secuencia a todos los evaluados. Un segundo tipo son *los tests adaptativos informatizados,* que permiten presentar los mejores ítems para cada evaluado. Por su importancia, dedicaremos a este tipo de tests un apartado propio.

En general, informatizar un test supone ciertas ventajas:

- ✔ Ayuda a estandarizar mejor las condiciones de aplicación de los tests para todos los evaluados: instrucciones comunes, control del tiempo de aplicación, reducción de la posibilidad de copia y de la transmisión del contenido de los tests, eficiencia en la corrección de respuestas, etc.
- ✔ Resulta necesario para la aplicación de los complejos procedimientos de estimación que se requieren en Teoría de la Respuesta al Ítem (TRI) (véase en este número Muñiz, 2010), con lo que ha permitido aplicar nuevos modelos psicométricos y hacer operativas sus eventuales ventajas.
- ✔ Permite proporcionar de forma inmediata información cuantitativa, verbal y gráfica sobre la posición de un evaluado respecto a un grupo en un baremo concreto, es decir, permite la elaboración de informes automáticos; es posible también proceder a una actualización continua de los baremos, incorporando a los mismos las puntuaciones de nuevos evaluados.
- ✔ El ordenador es necesario para aplicar *nuevos formatos de ítems* (por ejemplo presentaciones visuales dinámicas, ítems auditivos o secuencias simuladas grabadas en video), lo que ha representado una importante ampliación de los rasgos, competencias y comportamientos que pueden evaluarse en Psicología como, por ejemplo, la aptitud musical, el rendimiento de un controlador de tráfico aéreo, la capacidad para resolver conflictos, etc. (véase Drasgow y Olson-Buchanan, 1999). Se amplía así el rango de atributos que se pueden evaluar, aumentando la similaridad entre la tarea de evaluación y los criterios a predecir a partir de las puntuaciones en el test (por ejemplo las actividades a desempeñar por el evaluado en el puesto de trabajo). Además, puede romperse con el formato tradicional de respuesta (opción múltiple o categorías ordenadas) para plantear, por ejemplo, tareas tan diversas como marcar en un mapa determinadas localizaciones, seguir con el ratón el movimiento de un determinado objeto, rotar cier-

tos grados figuras tridimensionales, detectar y cambiar errores gramaticales de diversos textos, escribir con un editor de ecuaciones el resultado simplificado de una fórmula matemática, grabar una respuesta verbal en un micrófono, dar un diagnóstico médico después de recabar información diversa sobre los síntomas de un paciente o ubicar los componentes arquitectónicos de un edificio. En este tipo de ítems, además de los aciertos o errores, el ordenador permite registrar otro tipo de variables para medir el rendimiento (por ejemplo, los tiempos de reacción o las distancias físicas respecto a la solución óptima de una tarea visomotora).

- ✔ Algunos sistemas de evaluación informatizada permiten ya la corrección automática de la ejecución en una tarea concreta. En la figura 1, se muestra un ejemplo de un ítem de conocimientos sobre Botánica, que consiste en sombrear las zonas de distribución de una determinada especie y cuya corrección es automática (tomado de Conejo, Guzmán, Millán, Trella, Pérez de la Cruz y Ríos, 2004). Para puntuar este ítem, se usa un mapa con el sombreado correcto como plantilla. Si el estudiante marca aproximadamente la zona correcta (con un cierto margen de error) se puntúa como correcta la respuesta. Además se señala la proporción de área que es correctamente localizada. Por ejemplo: el 15.5% del área sombreada es correctamente sombreada y el 91.66% del área no sombreada es correctamente no sombreada.

El libro recientemente editado "*Automated Scoring of Complex Tasks in Computer Based Testing*" (Williamson, Mislevy y Bejar, 2006) recoge numerosos ejemplos de corrección automática en ítems de respuesta compleja. En este libro se aconseja elaborar *Diseños Centrados en la Evidencia* (DCE), en los que se plantea un esquema a seguir en este tipo de desarrollos. En la metodología DCE se parte de un modelo del evaluado (descripción exhaustiva de los constructos, habilidades o destrezas que queremos medir) y de un modelo de la tarea o familia de tareas (en el que se describen exhaustivamente las características de las tareas que permitirían generar el ítem de forma automática). El modelo de evidencia conecta ambos modelos recogiendo las relaciones entre el resultado del sujeto en la tarea y el constructo o la decisión sobre el evaluado (por ejemplo apto o no apto). En los DCE se diferencia entre reglas de evidencia o puntuación (que transforman el resultado del sujeto en las tareas en puntuaciones numéricas) y un modelo de medida (que conecta las puntuaciones numéricas con las puntuaciones en los constructos y con las decisiones que se tomen a partir de estos).

Uno de los primeros intentos importantes en el desarrollo de pruebas de corrección automática fue el ARE (Architectural Registration Examination), una batería de evaluación que desempeña un papel importante en el proceso de acreditación por el que en Canadá se otorgan licencias para ejercer como arquitecto. Algunos ítems exigen que el evaluado maneje algunas funcionalidades básicas de una herramienta informática para el diseño gráfico (ver figura 2). La tarea del evaluado es hacer un diseño de una vivienda, clínica... que cumpla con un conjunto de exigencias. Los diseños producidos

FIGURA 1

por el evaluado son puntuados automáticamente por un algoritmo atendiendo a la seguridad, la funcionalidad, la atención a las restricciones (geográficas, ambientales, climáticas…), la accesibilidad, etc. El desarrollo de estos procedimientos automáticos requiere de la colaboración de expertos y de la formación de grupos de discusión que permitan diseñar un algoritmo que proporcione puntuaciones similares a las que proporcionaría un evaluador humano. Paradójicamente, aunque los expertos proporcionan las reglas de puntuación incorporadas en el algoritmo, la corrección automática puede llegar a ser más eficiente, por una mayor sistematicidad en la aplicación de los criterios. El test ARE es un test de desempeño. El artículo de Martínez-Arias de este número los estudia en detalle.

FIGURA 2

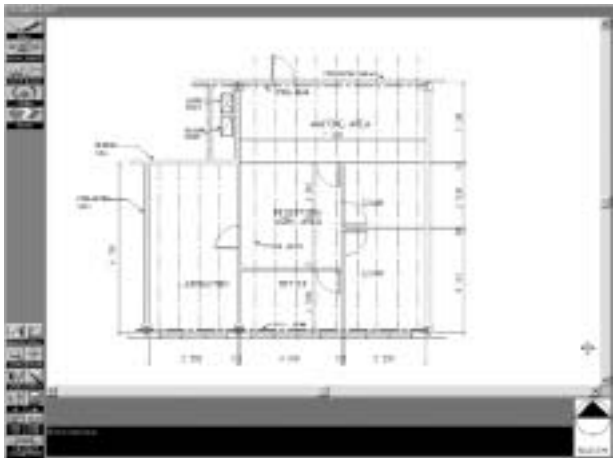
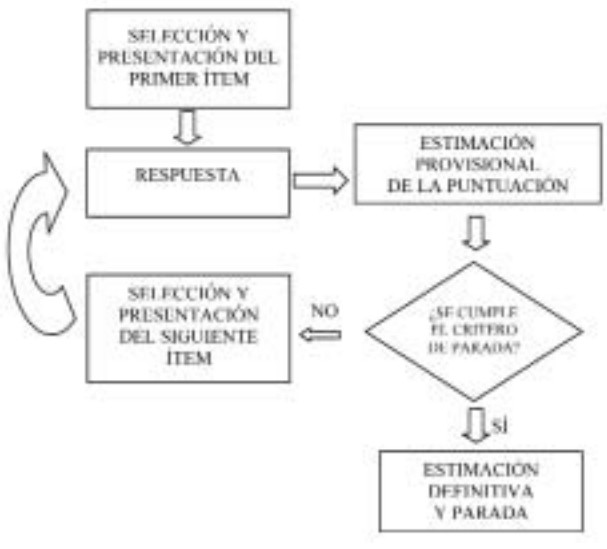


FIGURA 3



Tests adaptativos informatizados

El uso de los ordenadores combinado con la TRI permite la construcción de *tests adaptativos informatizados* (TAIs), cuya principal característica es que los ítems a administrar se van adaptando al nivel de competencia que va manifestando el evaluado, según sus respuestas a los ítems previos. Partiendo de un banco de ítems amplio, distintos ítems de ese banco son seleccionados para cada persona. Gracias a la TRI, las estimaciones del nivel de rasgo obtenidas en los distintos tests serán comparables (se encontrarán en la misma métrica).

La idea básica consiste en presentar únicamente los ítems que resultan altamente informativos para estimar el nivel de cada sujeto en un determinado rasgo. Una vez calibrado el banco de ítems, el proceso de aplicación de un TAI a un evaluando puede resumirse, de forma simplificada, en el diagrama de flujo de la Figura 3 (Olea y Ponsoda, 2003).

La aplicación de un TAI se inicia con una determinada estrategia de arranque, que consiste en establecer de alguna forma el nivel de rasgo inicial que se asigna al evaluando (por ejemplo, el nivel promedio de la población). Después de que el evaluando responde a cada ítem, se realiza una estimación de su nivel de rasgo mediante procedimientos estadísticos bayesianos o máximo-verosímiles. Se requiere también un algoritmo para la selección sucesiva de ítems. Generalmente, se emplean procedimientos basados en la medida de información, $I(\theta)$; por ejemplo, puede seleccionarse como segundo ítem el más informativo para el nivel θ estimado tras la primera respuesta. En contextos de acreditación, promoción o selección es importante que se muestreen los contenidos adecuadamente o que los evaluados reciban, en la medida de lo posible, ítems distintos. En esos casos, un algoritmo adecuado de selección deberá incluir restricciones en la tasa de exposición de los ítems (por ejemplo, que cada ítem no sea administrado en más del 20% de los tests) u otras restricciones, para que se garantice un adecuado muestreo de contenidos. Se requiere finalmente algún criterio para dar por terminada la secuencia de presentación de ítems, que normalmente tiene que ver con la consecución de cierto nivel de precisión o con haberse aplicado un número prefijado de ítems; esto último suele ser necesario para mantener el balance de contenidos en la prueba, y preferible para evitar en los usuarios del TAI la sensación de que se les

ha medido con pocos ítems. Como se representa en el diagrama, el ciclo "seleccionar ítem - aplicar ítem - recoger respuesta - estimar rasgo" se repite hasta que se satisface el criterio de parada.

Los TAIs, dada su condición adaptativa, tienen al menos tres importantes ventajas adicionales a las de cualquier test informatizado:

- ✔ Mejoran la seguridad del test, ya que gran parte de los ítems que se presentan a los evaluados son diferentes. Esta es una preocupación fundamental de los responsables de la evaluación en contextos aplicados porque, incluso cuando se decide aplicar tests convencionales, uno de los mayores obstáculos a la validez de los tests es que los evaluados puedan conocer de antemano los ítems que se les van a administrar.
- ✔ Reducen el tiempo de aplicación (a veces a menos de la mitad), ya que consiguen niveles similares de precisión que los tests convencionales con un número menor de ítems.
- ✔ Permiten además, con el mismo número de ítems que un test convencional, realizar estimaciones más precisas. Bajo condiciones similares a las de un test convencional (en tiempo requerido y número de ítems aplicados) un TAI permite mayores garantías (menor error de medida) respecto a los niveles de rasgo que se estiman y, por tanto, respecto a las decisiones que se toman a partir de las puntuaciones en los tests.

Estos tres aspectos resultan especialmente relevantes cuando se realizan aplicaciones masivas de tests de rendimiento o de conocimientos, por ejemplo en contextos de selección de personal, de evaluación educativa o en pruebas de certificación profesional o licenciatura. Por citar algunos ejemplos, en Estados Unidos existen versiones adaptativas informatizadas del TOEFL (para evaluar el nivel de inglés), del GRE (prueba de conocimientos para acceder a estudios universitarios), del GMAT (prueba de acceso a Escuelas de Negocios), del ASVAB (batería de aptitudes del Ejército) y de diversos exámenes de acreditación profesional (por ejemplo en Medicina y Enfermería) o de evaluación del nivel educativo de los estudiantes de Primaria y Secundaria. En España existen disponibles varios TAIs: el TRASI (Rubio y Santacreu, 2003) que mide la capacidad de razonamiento secuencial e inductivo; eCAT (Olea, Abad, Ponsoda y Ximénez, 2004) que mide el nivel de comprensión del inglés escrito; y CAT-Health (Rebollo, García-Cueto, Zardaín, Cuervo, Martínez, Alonso, Ferrer y Muñiz, 2009) para la evaluación de la calidad de vida relacionada con la salud. Se están elaborando otros para evaluar el dominio del catalán, euskera, otros idiomas, el ajuste emocional, la satisfacción con los servicios sanitarios, etc.

Aplicaciones vía web

La tecnología informática permite desde hace años su *aplicación a través de internet*. Por poner algunos ejemplos, se aplican a través de la web determinadas baterías neuropsicológicas, tests de conocimientos del idioma inglés, tests predictivos del rendimiento laboral, tests de conocimientos escolares, cuestionarios de personalidad aplicados en contextos clínicos o cuestionarios sobre adicciones a drogas (la información puede completarse en Bartram y Hambleton, 2006).

Tanto el test como los algoritmos de presentación y los resultados se almacenan y distribuyen desde un servidor, lo que permite un mayor control sobre los procesos de aplicación y una información inmediata sobre los resultados. La conexión a través de internet representa también importantes beneficios logísticos: una mayor accesibilidad a los evaluados (por ejemplo, en procesos de reclutamiento para la selección de personal o en casos de intervención psicológica de personas que residen lejos de los servicios de tratamiento) y, en algunos casos, un abaratamiento de costes (piénsese por ejemplo en la aplicación de tests a muestras numerosas de evaluados que viven en diferentes zonas geográficas de un país).

La aplicación a través de internet también supone ventajas para los editores de tests, ya que les permite tener acceso directo a bases de datos que permitan realizar los siempre necesarios estudios de validez de las puntuaciones y de "seguimiento" de las propiedades psicométricas de la prueba. Además, permite controlar que el "cliente" (por ejemplo, la empresa o institución que demanda la aplicación) tenga acceso únicamente a la información que resulte pertinente. Por ejemplo, ya no se requiere incluir plantillas de corrección, lo que implica una mayor garantía de seguridad.

Sin embargo, la utilización de internet como medio de transporte de los tests y de las respuestas de los evaluados requiere tener en cuenta algunas consideraciones en relación a varios riesgos:

- ✔ *Calidad.* Cualquiera puede acceder a centenares de tests que se ofrecen en todo el mundo y de los que desconocemos sus propiedades psicométricas. Como en otros muchos temas, un psicólogo competente debería saber filtrar bien los instrumentos de evaluación disponibles en la web que auténticamente han demostrado su utilidad, de aquellos que sirven únicamente como pasatiempos.

- ✔ *Seguridad.* Un importante problema es el de la seguridad del propio test, sobre todo cuando las puntuaciones en los tests tienen importantes consecuencias para los evaluados (admisión a un centro educativo, a un puesto de trabajo, acreditación profesional, etc.). En el caso del examen GRE, aplicado hace años vía internet, la empresa responsable de la prueba decidió volver a versiones de lápiz y papel tras comprobar la gran cantidad de ítems que los evaluados de ciertos países asiáticos conocían de antemano, debido a su transmisión en ciertos foros. Como es lógico, el acceso a los contenidos del test y a la información que proporcionan los evaluados debe ser seguro y controlado. A veces internet puede entrar en colisión con la Ley de Protección de Datos.
- ✔ *Control*. Otro problema importante tiene que ver con las posibilidades de suplantación de identidad, es decir, que sean otras personas las que respondan al test. Una posible solución sería la aplicación controlada por supervisores que aseguren la identidad de los evaluados, que asignen las contraseñas oportunas de acceso y que controlen el cumplimiento de las condiciones de aplicación.
- ✔ *Garantías tecnológicas.* La aplicación informatizada puede suponer una amenaza a la validez de las puntuaciones si las condiciones de evaluación no están estandarizadas. Por ejemplo, algunos tests que incluyen información dinámica y tiempos limitados de respuesta son muy susceptibles a la velocidad de transmisión de la información por la red y a las características del ordenador y conexión que tiene cada evaluado.

Por otro lado, conviene no olvidar que las propiedades de un test no dependen sólo de los ítems que se aplican sino también de cómo se aplican (por ejemplo, que el evaluador genere una situación adecuada de evaluación, que se responda a las dudas que surjan, que se garantice que el evaluado dedica el tiempo adecuado a las instrucciones, etc.). La necesidad de un supervisor directo de la aplicación puede depender del tipo de test (rendimiento óptimo vs. rendimiento típico) y de las consecuencias de la aplicación evaluación, entre otros aspectos.

Éstos y otros problemas han requerido la elaboración de directrices sobre buenas prácticas en el diseño y aplicación de tests informatizados, reservando recomendaciones específicas para los que se aplican a través de internet (ITC, 2005) y que plantean demandas adicionales en el control de calidad de este tipo de tests. Hay que determinar los requisitos mínimos de software y hardware, establecer mecanismos de prevención y detección de errores en la administración, prevenir y detectar brechas en la seguridad, determinar el nivel de supervisión en la aplicación, establecer controles de identificación del evaluado, garantizar el almacenamiento seguro de las respuestas, chequear periódicamente las propiedades psicométricas de los ítems, etc. En el ámbito más estrictamente psicométrico, las directrices establecen que un test informatizado debe incorporar la oportuna información psicométrica (fiabilidad y validez) y debe garantizar que no requiere otros conocimientos o destrezas (por ejemplo, la familiaridad con los ordenadores) diferentes a las que exige el test. Estas directrices se pueden consultar en la dirección de la ITC: http://www.intestcom.org/guidelines/index.php.

OTROS NUEVOS TIPOS DE TESTS

Tests basados en modelos

Un modo de obtener información sobre las inferencias que podemos realizar con las puntuaciones de un test es analizar los procesos, estrategias y estructuras de conocimiento que están implicados en la resolución de los ítems. Bejar (2002) emplea la denominación de *tests basados en modelos* para referirse al diseño de instrumentos de evaluación guiados por una teoría psicológica sobre el procesamiento de respuestas.

Un excelente muestrario de este tipo de tests se incluye en el libro de Irvine y Kyllonen (2002) *"Item generation for test development"* donde se recoge el progresivo acercamiento entre Psicología Cognitiva y Psicometría, lo que se ha traducido en la elaboración de tests de razonamiento cuantitativo, razonamiento analítico, visualización, analogías verbales, etc. El primer paso en la construcción de este tipo de pruebas es un análisis de los procesos cognitivos que demanda la resolución de la tarea y un estudio detallado de cuáles son las características del ítem que, en función de esos procesos, determinan su diferente nivel de demanda cognitiva y, por tanto, su dificultad. Por ejemplo, Hornke (2002) describe un test de rotación de figuras donde se manipulan variables como la cantidad de elementos a procesar, si las figuras son bi o tridimensionales, el ángulo de la rotación o el número y tipo de rotaciones (de derecha a izquierda, de arriba abajo...). Describe también un test de memoria visual donde los ítems son planos de una ciudad donde aparecen determinados iconos para representar ciertos servicios públicos, manipulándose en cada caso la cantidad de iconos, su tamaño o el nivel dispersión en el mapa.

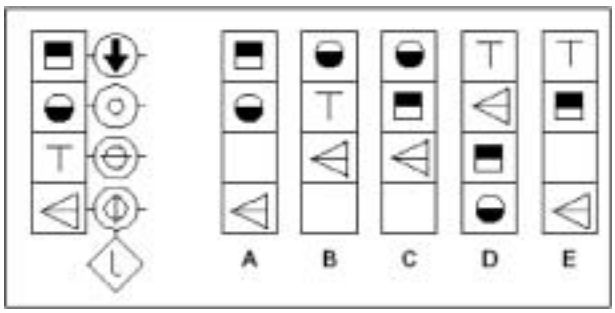
En España, Revuelta y Ponsoda (1998) desarrollaron un test basado en un modelo cognitivo para el test DA5. Los 50 ítems del test pretenden medir la capacidad de razona-

miento lógico mediante tareas que incluyen un conjunto de instrucciones (símbolos dentro de los círculos y del rombo) sobre lo que debe hacerse mentalmente con la figura adyacente correspondiente (véase figura 4). Un ítem consta de varias figuras (columna de cuatro cuadrados a la izquierda de la Figura 4, que contiene cada uno una figura), las instrucciones de los cambios que se han de hacer con cada figura (columna de círculos y rombo), y de las cinco posibles respuestas (columnas A, B … E). La tarea del evaluado es aplicar a las figuras las instrucciones y elegir la opción correcta de las cinco posibles. Las instrucciones pueden requerir, por ejemplo, girar la figura cierto número de grados, intercambiar la posición con la figura anterior, omitirla, ignorar otras instrucciones o reordenar de determinada forma todas las figuras.

Un modelo de procesamiento asume que el evaluado codifica la primera figura (la que aparece en el primer cuadrado de la primera columna) y la instrucción, aplica la instrucción sobre la figura (en ejemplo, la instrucción indica que la figura ha de desplazarse un cuadrado hacia abajo, por lo que sólo las opciones C y E podrían ser correctas), y sigue secuencialmente con las demás figuras hasta alcanzar la solución. Se estudió la aportación de cada una de las instrucciones (y de las veces que es necesario aplicarlas) a la dificultad de los ítems, mostrando más peso en la predicción de la dificultad las instrucciones que requerían reordenar las 4 figuras mentalmente.

Una aportación novedosa de esta manera de proceder es que si conocemos las variables que intervienen en los procesos de respuesta, puede establecerse un método para construir todo el universo posible de ítems gobernado por dichas variables. El procedimiento, denominado "*generación automática de ítems*" (GAI), consiste en la construcción de bancos de ítems mediante algoritmos. En la GAI se establece un conjunto de reglas explícitas, susceptibles de programarse en un ordenador, que determinan cómo deben construirse los ítems. Por ejemplo,

FIGURA 4



Revuelta y Ponsoda (1998) generaron los 4.242 ítems posibles que tienen su base en el DA5, combinando el tipo de figuras, las instrucciones a aplicar y determinados criterios para generar opciones incorrectas de respuesta. Si el modelo que describe los procesos de respuesta de los ítems es correcto, resultará posible conocer la dificultad de nuevos ítems antes de que hayan sido aplicados a persona alguna. Son muy importantes las ventajas de disponer de todo el banco posible de ítems, principalmente para garantizar que se mide con elevada precisión cualquier nivel de capacidad.

Tests ipsativos

Fundamentalmente en contextos de selección de personal, el falseamiento de respuestas a los tests de personalidad es un problema que se ha intentado resolver de varias formas. Una de las más alentadoras es la elaboración de tests ipsativos, que obligan al evaluado a elegir entre opciones de respuesta que tienen un nivel similar de deseabilidad y que se refieren a dimensiones diferentes de personalidad. Por ejemplo, el aspirante puede tener que elegir entre *"soy una persona trabajadora"* (responsabilidad) y *"soy una persona abierta"* (extraversión). El proceso de diseño de un test ipsativo es básicamente el siguiente:

a. Determinar las dimensiones a evaluar y los ítems iniciales que las definen.
b. Diseñar con estos ítems iniciales un test normativo convencional. Conviene realizar estudios factoriales para determinar empíricamente los ítems que forman cada dimensión y, en su caso, eliminar los ítems que no saturan en el factor previsto.
c. Establecer el número de opciones de cada ítem ipsativo. Lo más simple es establecer ítems binarios, cada uno formado por dos ítems iniciales.
d. Realizar un estudio empírico donde una muestra apropiada de jueces valore el nivel de deseabilidad de cada ítem inicial. A partir de estos juicios se obtienen valores en deseabilidad para cada uno de los ítems iniciales.
e. Diseñar el test ipsativo, considerando que en los ítems deben todas las posibles combinaciones de dimensiones. En cada ítem ipsativo deben incluirse opciones (ítems iniciales) de similar deseabilidad. Cada dimensión se debe comparar con cualquier otra un número similar de veces.
f. Establecer el sistema de puntuación de los evaluados, por ejemplo, contando las veces que eligen las opciones de cada una de las dimensiones.

Ejemplo del proceso de elaboración de un test ipsativo
(Abad, Olea, Ponsoda y Garrido, 2007)

1) Dimensiones a evaluar: las 5 dimensiones de personalidad definidas en el modelo Big-Five, cada una evaluada mediante 18 adjetivos.
2) Test normativo. aplicación de los 90 ítems a una muestra según un formato de 5 categorías ordenadas, pidiendo el grado en que cada uno le describe a la persona.
3) Estudio factorial: se retuvieron los 12 ítems de cada dimensión que mayor saturación manifestaron en el factor previsto, con lo que el test definitivo constaba de 60 ítems.
4) Obtención de índices de deseabilidad (ID). una muestra de personas valoró (de 1 a 4) el grado en que cada adjetivo indicaba una cualidad positiva para ser eficiente en un determinado puesto laboral. Las medias de estas valoraciones se consideraron como índices de deseabilidad de los ítems. El adjetivo de menor media fue "corriente" (ID = 1,93) y el de mayor media "organizado" (ID =3,87).
5) Diseño del test ipsativo. Se decidió construir un test de 30 ítems ipsativos, cada uno formado por dos adjetivos de dimensiones distintas y similar ID. Por ejemplo, uno de los ítems era "estable-enérgico" que, respectivamente, se refieren a las dimensiones de estabilidad emocional y extraversión, y que obtuvieron valores en ID de 3,71 y 3,43. Según este diseño, cada dimensión se comparaba 3 veces con las otras 4 dimensiones de personalidad restantes.
6) Puntuación en el test ipsativo. Para puntuar a cada sujeto en cada una de las 5 dimensiones, se sumaron las veces que en los pares de adjetivos se elegían los ítems de cada una de ellas. Por tanto, la puntuación máxima teórica en una dimensión fue 12, mientras que la mínima 0.
7) Se realizaron estudios de validez convergente y predictiva (correlaciones con calificaciones en cursos de formación), mostrando la mayor capacidad predictiva algunos ítems ipsativos que combinaban adjetivos de las dimensiones de estabilidad emocional y responsabilidad.

En las últimas décadas, los tests ipsativos han tenido momentos de auge y declive, con defensores y detractores que con igual fuerza argumentan sus beneficios o problemas. Algunos de estos problemas son:

a. El modo de puntuar ipsativamente a un sujeto en las diferentes dimensiones provoca interdepencias entre éstas: una puntuación muy alta en una dimensión necesariamente conlleva puntuaciones no elevadas en las restantes. Este problema es tanto mayor cuanto menor el número de dimensiones. De forma más general, el promedio de las correlaciones entre *m* dimensiones se acerca a *-1/(m-1*), siendo *m* el número de dimensiones (Meade, 2004). En el caso de medir dos únicas dimensiones, la correlación entre ambas sería necesariamente -1. El modo ipsativo de puntuación lleva además a que sea cero la suma de las covarianzas de las dimensiones con una variable externa (por ejemplo, un criterio) y a distorsiones en los coeficientes de fiabilidad para las puntuaciones en las dimensiones. Todo esto exige un tratamiento psicométrico específico de los datos ipsativos (no es raro, por ejemplo, que las soluciones factoriales de datos normativos e ipsativos del mismo test sean diferentes), que actualmente es objeto de investigación.
b. Conceptualmente, un test ipsativo plantea una tarea de preferencias y, por tanto, permite la comparación entre escalas dentro de una persona (por ejemplo, podría decirse que una persona es más responsable que extravertida) pero no entre distintas personas (que una persona sea más responsable que otra). Por ello, su uso está más indicado en las medidas de atributos que impliquen preferencias, como es usual en la medición de los intereses.
c. No es claro que sean resistentes al falseamiento ya que los aspirantes pueden ser conscientes de cuáles son las dimensiones deseables para el puesto.

No nos parece muy recomendable por el momento la aplicación de tests ipsativos si se pretende realizar comparaciones de rendimiento entre diferentes evaluados, dado que resulta complicado estudiar sus propiedades psicométricas mediante modelos y técnicas usuales. Sin embargo, vemos una importante potencialidad a este tipo de tests (algunos estudios han mostrado ya una mayor validez predictiva que los tests usuales de personalidad) cuando se consoliden algunos intentos que se están realizando en el ámbito de la investigación psicométrica para modelar teóricamente las respuestas a este tipo de ítems (Stark, Chernyshenko y Drasgow, 2005). En cualquier caso, la cuestión está lejos de ser resuelta.

Tests conductuales

En el contexto de la medición de la personalidad, existe una línea teórica de evaluación comportamental de la personalidad donde se estudian los estilos interactivos o tendencias de comportamiento constantes ante situaciones determinadas (Santacreu, Rubio y Hernández, 2006). Desde esta perspectiva se diseñan tests comportamentales informatizados para medir, por ejemplo, la tendencia al riesgo (propensión a elegir las opciones más recompensadas a pesar de ser poco probables) mediante simulaciones de juegos de ruleta o dados, o mediante tareas de toma de decisiones más o menos proclives a accidentes. En la figura 5 se muestra tarea consistente en decidir cuándo cruzar la calle para ir lo más rápido posible a una farmacia, cambiando en los sucesivos ensayos la ubicación de la persona y sabiendo que puede aparecer un coche del túnel. Si el peatón se

encuentra muy a la izquierda aumenta la probabilidad de que sea atropellado (menor visibilidad) pero también se reduce el tiempo para llegar a la farmacia. Lo más seguro es moverse hacia la derecha y cruzar, pero eso conlleva un mayor tiempo. Tras cada ensayo, el evaluado recibe feedback sobre el tiempo que ha tardado en llegar pero no sobre si ha sido atropellado. La tendencia al riesgo se obtiene calculando la media en los sucesivos ensayos de la distancia entre la persona y la farmacia (mayor media, menor tendencia al riesgo). Obviamente, este modo de proceder es muy distinto a aplicar tests de personalidad donde las personas informan de su tendencia a la búsqueda de sensaciones o su nivel de apertura, tal como se hace en los tradicionales tests de personalidad. Los profesionales que optan por este tipo de tests consideran que una de sus ventajas tiene que ver con la eliminación de los problemas de deseabilidad.

Tests situacionales

Consisten en describir ciertas situaciones (por ejemplo, en el ámbito laboral) y pedir a los sujetos que digan cómo creen que reaccionarían ante dichas situaciones. Parece que este tipo de pruebas añaden poder predictivo de la eficacia laboral a los tradicionales tests de capacidad cognitiva y de personalidad (por eso se aplican cada vez más frecuentemente), aunque son escasos los estudios que se han realizado sobre su eficacia para reducir el falseamiento de respuestas. Pueden realizar descripciones en un formato de respuesta abierta o, lo que es más usual, elegir entre varias posibilidades que se describen de antemano. A continuación se presenta un ejemplo de un ítem situacional sobre integridad (Becker, 2005). Entre corchetes se especifica el modo de puntuación de las respuestas, establecido a partir de las opiniones de expertos:

Tu equipo de trabajo está en una reunión debatiendo sobre cómo vender un producto nuevo. Todos parecen estar de acuerdo en que el producto debe ser ofertado a los clientes en el presente mes. Tu jefe tiene mucho interés en que sea así, y tú sabes que a él no le gustan los desacuerdos en público. Sin embargo, tú tienes reservas porque un informe reciente del departamento de investigación apunta hacia diversos problemas potenciales de seguridad. ¿Cuál crees que sería tu reacción?

A. Tratar de comprender por qué todos los demás quieren ofertar el producto a los clientes en este mes. Tal vez tus preocupaciones están fueran de lugar. [-1]

B. Expresar tus preocupaciones con el producto y explicar por qué crees que las cuestiones de seguridad necesitan ser abordadas. [1]

C. Mostrarte de acuerdo con lo que los demás quieren hacer para que todos se sientan bien acerca del equipo. [-1]

D. Después de la reunión, hablar con algunos de los otros miembros del equipo para ver si ellos comparten tus preocupaciones. [0]

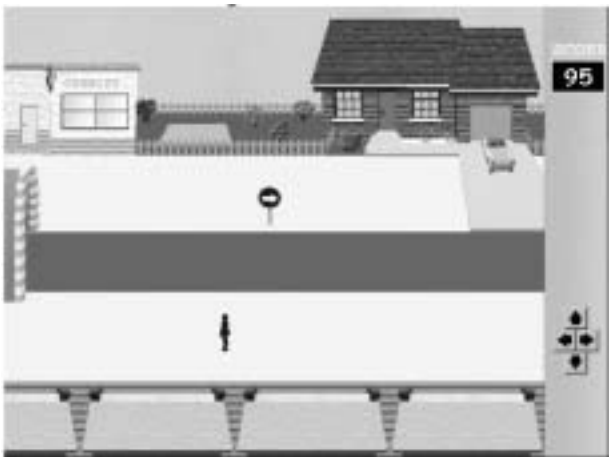
Desde un punto de vista psicométrico, un tema especialmente relevante es cómo puntuar de la mejor forma las respuestas a este tipo de ítems. Bergman, Donovan, Drasgow, Henning y Juraska (2006) estudiaron los diferentes efectos que tienen 11 modos diferentes de puntuar los ítems de un test situacional para evaluar la capacidad de liderazgo, formado por 21 ítems que se presentan mediante video y que tienen cuatro opciones diferentes de respuesta según el grado de participación en la toma de decisiones.

ALGUNOS RIESGOS ADICIONALES, ALGUNOS RECURSOS

Las importantes ventajas que tienen para el psicólogo los nuevos tipos de tests no pueden quedar empañadas por ciertos riesgos que queremos subrayar.

En primer lugar, debe enfatizarse que las nuevas tecnologías no son "per se" garantes de mejores mediciones. La eficiencia de los nuevos procedimientos de respuesta y procesamiento de la información no puede sustituir al necesario escrutinio psicométrico de las puntuaciones asignadas. La apariencia de validez de los nuevos formatos de ítems debe ir acompañada de evidencias empíricas de validez. Por ejemplo, un ítem "multimedia" puede ser más informativo que un ítem clásico de opción múltiple pero puede requerir mucho más tiempo para su resolución. Por otro lado, las demandas de precisión requerida pueden ser distintas si el objetivo es clasificar a una persona (apto vs. no apto) o si el objetivo es cuantificar su nivel de rasgo. También es necesario reflexionar sobre cuándo merece la pena aplicar un TAI y cuándo no (Wainer, 2000). Por ejemplo, si el test

FIGURA 5



es de altas consecuencias para los evaluados, se va a aplicar una o dos veces al año y el contenido no requiere una aplicación informatizada, los costes de un TAI (necesidad de crear y mantener grandes bancos de ítems, de desarrollar, evaluar y actualizar el software, disponibilidad de ordenadores para la aplicación, etc.) pueden superar a los beneficios.

Un segundo riesgo tiene que ver con el olvido de ciertas áreas de aplicación de los tests. Desarrollar nuevos tipos de tests es costoso y se corre el riesgo de avanzar casi exclusivamente en contextos aplicados (organizacionales o educativos) donde más recursos económicos se invierten o donde más se precisan soluciones tecnológicas eficientes. Los avances no deben olvidar determinados contextos de medición estrictamente propios de nuestra profesión, como son la evaluación clínica o la evaluación de programas de intervención psicosocial. En este sentido, la conjunción entre los nuevos tipos de tests, los nuevos modelos psicométricos y los modelos estadísticos para medir el cambio conseguido por las intervenciones debería ser un terreno fructífero para proyectos de I+D.

Un tercer riesgo se refiere al mal uso de los nuevos tests. Debido a su inmediata disponibilidad, los nuevos tipos de tests pueden aplicarse en contextos inadecuados, por personas no preparadas y realizando inferencias erróneas a partir de las puntuaciones que proporcionan.

¿Qué puede hacer el profesional de la Psicología para incrementar su competencia en los nuevos modos de medir? Más que nunca se necesita una formación continua a lo largo de la vida profesional para estar al tanto de las innovaciones que aceleradamente se van produciendo para mejorar la medición psicológica. Aún resultando ciertamente atrevido dar consejos, el profesional competente podría comenzar leyendo alguno de los libros recientes de Psicometría y las revistas especializadas donde se publican los avances psicométricos y las experiencias aplicadas con nuevos tipos de tests (algunas de las españolas más sensibles a estos temas son *Psicothema*, *Psicológica, Revista Electrónica de Metodología Aplicada* o *Spanish Journal of Psychology*). La revista *Psicológica* publicó en el 2000 un número monográfico sobre TAIs. Información en castellano sobre los TAIs puede encontrarse en libros (Olea y Ponsoda, 2003; Olea et al., 1998) y capítulos de libros (Olea y Ponsoda, 1996). En inglés, son muchos los libros sobre tests informatizados (Bartram y Hambleton, 2006; Eggen, 2004; Mills, Potenza, Fremer y Ward, 2002; Parshall, Spray, Kalohn y Davey, 2002; Sands, Waters y McBride, 1997; van der Linden y Glas, 2000; Wainer, Dorans y col., 2000).

Un sencillo tutorial sobre los TAIs puede encontrarse en la dirección http://edres.org/scripts/cat/catdemo.htm y una página fundamental para el investigador, que recoge una amplia información teórica y aplicada sobre los TAIs, es la siguiente: http://www.psych.umn.edu/psylabs/catcentral/.

Puede consultarse también los catálogos de tests disponibles en la web de las principales empresas editoras. Los profesionales interesados en mejorar su formación sobre estos temas pueden asistir a cursos concretos sobre estos temas. Nuestras universidades ofrecen varios. Puede también empezar a manejar algunos programas disponibles para la elaboración y análisis de tests informatizados. Si esta es su opción, no deje de consultar las prestaciones que se ofrecen en la siguiente dirección de la universidad de Málaga (http://jupiter.lcc.uma.es/siette.wiki.es/index.php/Portada) o en la principal distribuidora norteamericana de software psicométrico: http://assess.com. Además del software general para aplicar la TRI, existe software para implementar TAIs (FASTEST y POSTSIM 2.0); mientras que el primero (ASC, 2001) permite la organización de bancos de ítems, ensamblaje de pruebas y aplicación de TAIs, el segundo permite evaluar el funcionamiento psicométrico de un TAI mediante simulación y bajo distintas condiciones (de selección de ítems, de estimación del nivel de habilidad y de criterio de parada). Para tener un "primer contacto" también puede servir el programa ADTEST (Ponsoda, Olea y Revuelta, 1994).

REFERENCIAS

Abad, F.J., Olea, J., Ponsoda, V. y Garrido, L. (2007). *Test POLIPSA: Informe técnico y propiedades psicométricas.*

ASC (2001). *The FastTEST Professional Testing System, Version 1.6.* [*Computer software*]. St. Paul, MN: Author.

Bartram, D. y Hambleton, R. K. (2006). *Computer-based testing and the internet issues and advances*. Chichester, West Susex: Wiley.

Becker, T. E. (2005). Development and validation of a situational judgment test of employment integrity. International Journal of Selection and Assessment, 13(3), 225-232.

Becker TE. (2005). Development and validation of a situational judgment test of employee integrity. *International Journal of Selection and Assessment, 13,* 225–232.

Bejar, I.I. (2002). Generative testing: From conception to implementation. En S. H. Irvine y P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-217). Mahwah, NJ: LEA.

Bergman, M.E., Drasgow, F., Donovan, M.A., y Henning, J.B. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, *14*, 223-235.

Conejo, R., Guzmán, E., Millán, Trella, M., Pérez-de-la-Cruz, L. y Rios, A., (2004): SIETTE: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education*, 14, 29-61.

Davey, T. (2005). Computer-based testing, En B. S. Everitt y D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science*. Hoboken, NJ: Wiley.

Drasgow, F. y Olson-Buchanan, J. (1999). *Innovations in computerized assessment*. Mahwah, NJ: LEA.

Eggen, T. J. H. M. (2004). *Contributions to the theory and practice of computerized adaptive testing.* Arnhem, Holanda: Citogroep.

Hornke, L. F. (2002). Item-generative models for higher order cognitive functions. En S. H. Irvine y P. C. Kyllonen (Eds.), Item generation for test development. New Jersey: Lawrence Erlbaum Associates.

International Test Comission (2005). *International Guidelines on Computer-Based and Internet Delivered Testing.* Recuperado el 30 de junio de 2005, http://www.intestcom.org.

Irvine, H. y Kyllonen, P.C. (2002), *Item generation for test development*. New Jersey: LEA.

Meade, A. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*, 531-552.

Mills, C. N., Potenza, M. T., Fremer, J. J. y Ward, W. C. (2002). *Computer-based testing: Building the foundation for future assessments.* Mahwah, NJ: LEA

Muñiz, J. (2010). Las teorías de los tests: Teoría Clásica y Teoría de Respuesta a los Ítems. *Papeles del Psicólogo, 31(1)*, 57-66.

Muñiz, J. y Fernández-Hermida, J.R. (2010) La Opinión de los Psicólogos Españoles sobre el Uso de los Tests. *Papeles del Psicólogo, 31(1)*, 108-122.

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

Olea, J. y Ponsoda, V. (2003). *Tests adaptativos informatizados*. Madrid: UNED.

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

Olea, J., Abad, F. J., Ponsoda, V. y Ximenez, M. C. (2004). A computerized adaptive test for the assessment of written English: Design and psychometric properties. *Psicothema*, *16*, 519-525.

Parshall, C. G., Spray, J. A., Kalohn, J. C., y Davey, T. (2002). *Practical considerations in computer-based testing*. Nueva York: Springer.

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

Quintana, J., Bitaubé, A. y López-Martín, S. (2008). *El lugar de la Psicología en la universidad española del siglo XX*. UAM Ediciones: Madrid.

Rebollo, P., García-Cueto, E., Zardaín, J.C., Martínez, I., Alonso, J., Ferrer, M. y Muñiz, J. (2009). Desarrollo del CAT-Health, primer test adaptativo informatizado para la evaluación de la calidad de vida relacionada con la salud en España. *Medicina Clínica*, 133, 7, 241-251.

Revuelta, J. y Ponsoda, V. (1998). Un test adaptativo informatizado de análisis lógico basado en la generación automática de ítems. *Psicothema*, *10*, 3, 709-716.

Rubio, V. y Santacreu, J. (2003). *TRASI Test adaptativo informatizado para la evaluación del razonamiento secuencial y la inducción como factores de la habilidad intelectual general.* Madrid: TEA ediciones.

Sands, W. A., Waters, B. K. y McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association.

Santacreu, J., Rubio, V.J., y Hernández, J.M. (2006). The objective assessment of personality: Cattell's T-data revisited and more. *Psychology Science*, 48, 53-68.

Stark, S, Chernyshenko, O.S., Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multiunidimensional pairwise-preference model. *Applied Psychological Measurement*, *29*, 184–203.

van der Linden, W. J. y Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. Londres: Kluwer Academic.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B., Mislevy, R., Steinberg, L. y Thissen, D. (2000). *Computerized adaptive testing: A primer (2ª ed.)*. Mahwah, NJ: LEA.

Wainer, H. (2000). CAT: Wether and Whence. *Psicologica*, *21*, 121-133.

Williamson, D. M., Mislevy, R. J. y Bejar, I. I. (2006). *Automated Scoring of Complex Tasks in Computer Based Testing.* Mahwah, NJ: LEA.