anales de psicología
2012, vol. 28, nº 1 (enero), 289-302

ISSN edición impresa: 0212-9728. ISSN edición web (http://revistas.um.es/analesps): 1695-2294

Tests adaptativos informatizados: una perspectiva general

Juan Ramón Barrada

Universidad de Zaragoza

Resumen: Los tests adaptativos informatizados (TAIs) adaptan las preguntas a administrar a cada examinado según sus respuestas a las preguntas previas. De este modo, se consiguen estimaciones de su nivel de rasgo más precisas o se reduce la longitud del test. En los últimos años, se han desarrollo diversos TAIs en España y es probable que, dadas las ventajas que ofrece esta técnica, sean bastantes más lo que se hagan disponibles próximamente. El objetivo de este trabajo es ofrecer una visión actualizada de este campo. Para ello, se presenta la estructura básica de un TAI y se comentan los distintos pasos que lo componen. Se hace especial énfasis en la selección de ítems, la parte fundamental para la adaptatibilidad del test, desde la perspectiva de los cuatros objetivos que ha de satisfacer un TAI: (a) precisión, (b) seguridad del banco de ítems, (c) control de contenidos, y (d) mantenimiento de la prueba.
Palabras clave: tests adaptativos informatizados; fiabilidad; reglas de selección de ítems; seguridad del test.

Title: Computerized adaptive testing: a general perspective.
Abstract: Computerized adaptive testing (CAT) adapts the items to be administered to each examinee according to the responses to the previous items. In this way, more accurate trail level estimations can be obtained or test length is reduced. In the last years, several CATs have been developed in Spain and it can be expected that, given the advantages of this technique, more will become available soon. The goal of this work is to offer and updated view of this topic. For doing so, the basic structure of a CAT is presented and the different steps composing it are commented. Special emphasis is given ot item selection, the fundamental part for the adaptability of the test, from the perspective of the four objectives that must be satisfied by a CAT: (a) accuracy, (b) item bank security; (c) content balance; and (d) test maintenance.
Keywords: computerized adaptive testing; accuracy; item selection rules; test security.

Introducción

La Psicometría indica que los ítems varían en su calidad para la medición y que no todos los ítems son igualmente adecuados para todos los niveles de rasgo. Aquellos ítems que más estrechamente se vinculan con aquello que estamos midiendo presentan una mayor correlación con el total de la escala y, al ser calibrados según algunos modelos de Teoría de Respuesta al Ítem (TRI – Embretson y Reise, 2000; Muñiz, 1997), un mayor parámetro de discriminación. Aquellos ítems que mejor permitan evaluar a un examinado tendrán, en general, un parámetro de localización próximo al nivel de rasgo del examinado.

Teniendo esto en cuenta, en el caso de disponer de un banco de ítems marcadamente más amplio que la longitud del test que deseamos administrar, resultaría conveniente ajustar los ítems a presentar a cada examinado. En lugar de administrar tests fijos, en los que todos los evaluados reciben idénticos ítems, podríamos seleccionar aquellas preguntas que en mayor medida reducen la incertidumbre sobre el nivel de rasgo de cada examinado. La selección del ítem se haría teniendo en cuenta tanto los parámetros de los ítems como el nivel de rasgo del examinado. Las puntuaciones de los examinados que han respondido diferentes preguntas se situarían en la misma escala mediente métodos derivados de la TRI.

El nivel de rasgo de los examinados, una variable latente, es una información que resulta inaccesible. De lo que podemos disponer es de estimaciones que esperamos que sean progresivamente más precisas según aumenta el número de ítems administrados. Previamente a la administración de ningún ítem, la estimación que minimiza el error es asignar a un examinado el nivel de rasgo promedio en la población. Una vez administrados uno o más ítems, la estimación se realizará según el patrón de respuestas a los ítems contestados y los parámetros de éstos.

Estas ideas básicas son las que fundamentan los Tests Adaptativos Informatizados (TAIs – Olea y Ponsoda, 2001; van der Linden y Glas, 2000). Los TAIs, cuando se comparan con tests fijos, ofrecen varias opciones atractivas: (a) resulta posible igualar la precisión de medida para los diferentes niveles de rasgo; (b) resulta posible mantener la precisión reduciendo a aproximadamente la mitad el número de ítems administrados; o (c) en el caso de mantener la longitud del test fijo, se obtienen estimaciones más precisas. Las primeras propuestas teóricas son de los años setenta (p. ej., Lord, 1977; Urry, 1977), pero es en los años ochenta y noventa cuando empieza a popularizarse este modo de administración de tests. A finales del pasado siglo, ya se admistraban más de un millón de TAIs por año (Wainer, 2000a). Actualmente, importantes pruebas como el ASVAB (*Armed Services Vocational Aptitude Battery*), el TOEFL (*Test of English as a Foreign Language*) o el GMAT (*Graduate Management Admission Test*), entre otras, son administradas de un modo adaptativo. En España, los primeros TAIs comercializados llegan años más tarde: el TRASI (Rubio y Santacreu, 2004), que mide la capacidad de razonamiento secuencial e inductivo, y eCAT (Abad, Olea, Aguado, Ponsoda y Barrada, 2010; Olea, Abad, Ponsoda y Ximénez, 2004), que mide el nivel de comprensión del inglés escrito. Recientemente, se ha desarrollado un TAI desde el ámbito médico, CAT-Health (Rebollo et al., 2009), para la evaluación de la calidad de vida relacionada con la salud. Hasta donde sabemos, el último TAI español presentado evalúa el conocimiento de lengua vasca (López-Cuadrado, Pérez, Vadillo y Gutiérrez, 2010).

Hoy día, el campo de los TAIs, tanto a nivel de investigación como aplicado, y tanto a nivel nacional como interna-

Dirección para correspondencia [Correspondence address]: Juan Ramón Barrada. Facultad de Ciencias Sociales y Humanas. Universidad de Zaragoza. 44003 Teuel (España).
E-mail: juanramon.barrada@unizar.es

cional, es un área activa y en desarrollo. El propósito de este artículo es ofrecer una revisión actualizada y en castellano sobre el campo, en la que expondremos aquellos problemas más candentes en el área de los tests adaptativos y las propuestas más prometedoras para solucionarlos. Para ello, a continuación procederemos a: (a) ilustrar la estructura básica de un TAI; (b) definir cuáles son los objetivos que ha de satisfacer; y (c) desarrollar los modos de selección de ítems empleados en los TAIs, puesto que éste es el elemento central de este modo de administrar tests.

Estructura de un TAI

La Figura 1 recoge el diagrama de flujo de un TAI. En el arranque se inicializa el sistema, leyendo los enunciados y los parámetros de los ítems y demás requisitos del procedimiento informático. En la fase de fin del TAI se salvaría la información pertinente, se ofrecería la retroalimentación e instrucciones que correspondieran al examinado, etc. El cuerpo central del TAI lo compone un proceso iterativo con cuatro pasos. Ofrecemos una descripción de las principales propuestas para aplicar cada uno de ellos.

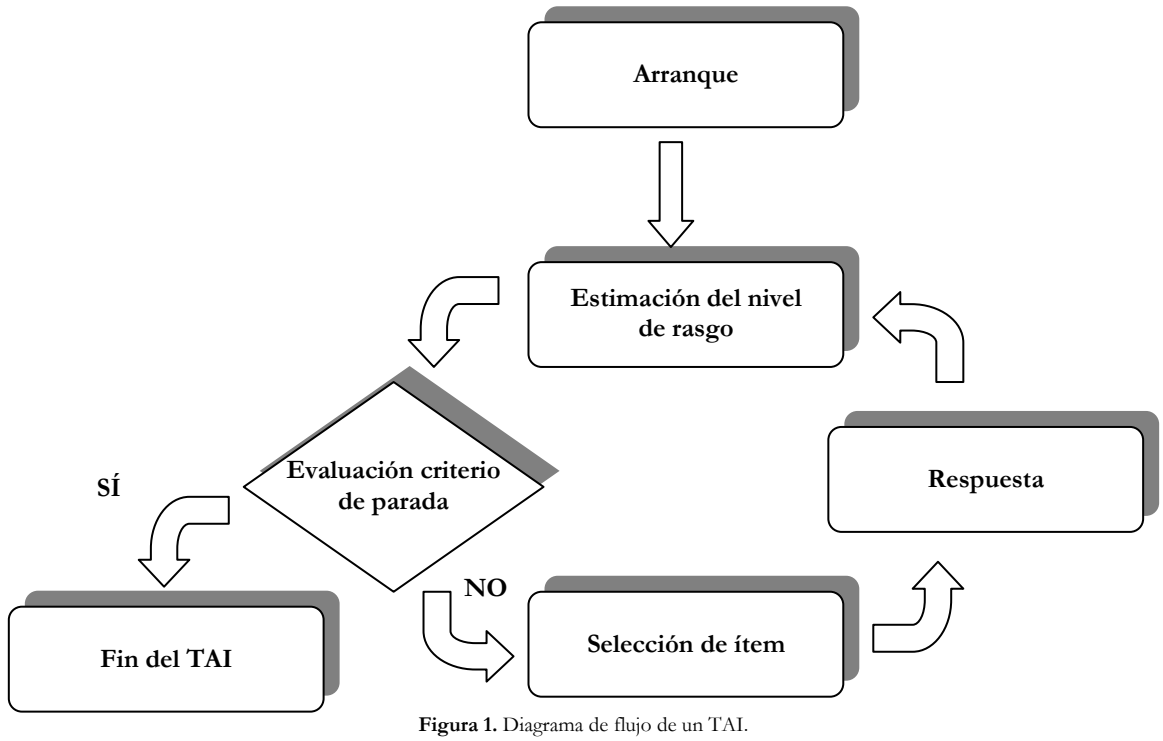

Figura 1. Diagrama de flujo de un TAI.

Estimación del nivel de rasgo

Hay que distinguir entre dos momentos diferentes en un TAI.

- El primero, cuando el examinado todavía no ha respondido a ítem alguno. En este caso, tres son las opciones básicas que se presentan (Parshall, Spray, Kalohn y Davey, 2002): (a) asignar como nivel de rasgo el promedio poblacional; (b) para incrementar la motivación de los examinados, asignar como nivel inicial un valor por debajo del promedio para aumentar la probabilidad de acierto, o (c) mediante información sobre los examinados que permita predecir sus desempeños en el test (p. ej., puntuaciones en otro test o nivel educativo) estimar un nivel de rasgo inicial diferente por examinado (van der Linden, 1999).
- El segundo, cuando el examinado ya ha respondido a alguna pregunta. En este caso, se aplican los métodos de estimación por máxima-verosimilitud o bayesianos para estimar el nivel de rasgo. Un supuesto básico en la mayor parte de los modelos de TRI es la independencia local. Asumiendo ésta, la probabilidad del patrón de respuestas observado es igual al producto de las probabilidades de la respuesta obtenidas en cada ítem. La función de verosimilitud es la que relaciona este producto con el continuo de niveles de rasgo:

$$L\left(\theta \middle| u_1 \ldots u_n, x_1 \ldots x_n\right) = \prod_{i=1}^{n} \left\{ P_i(\theta)^{u_i} \left(1 - P_i(\theta)\right)^{1-u_i} \right\}^{x_i}, \quad (1)$$

donde θ es el nivel de rasgo, $P_i(\theta)$ es la probabilidad de respuesta correcta del ítem *i*, *n* es el tamaño del banco de ítems y x_i es el indicador de administración / no administración (1 / 0) y u_i es el indicador de respuesta correcta / no correcta (1 / 0), ambos del ítem *i*.

El método de máxima-verosimilitud (Birnbaum, 1968) ofrece como nivel de rasgo estimado aquel en el que la función de verosimilitud encuentra su máximo:

$$\hat{\theta}_{q-1} = \arg\max_{\theta} L\left(\theta | u_1 \ldots u_n, x_1 \ldots x_n\right), \quad (2)$$

donde θ_{q-1} es el nivel de rasgo estimado después de los *q-1* primeros ítems, siendo *q* el indicador de la posición serial dentro del test del siguiente ítem a administrar.

Cuando se opta por la estimación máximo-verosímil, hay que tener en cuenta que la función de verosimilitud no tiene máximo en los números reales mientras el patrón de respuestas es constante, todas las respuestas aciertos o todas errores. Por eso, hasta que el patrón deja de ser constante, las alternativas son: (a) aplicar temporalmente una estimación bayesiana; (b) fijar el rango de valores admisibles de niveles de rasgo y permitir que la estimación se sitúe en uno de los extremos; o (c) utilizar un método por escalera (Dodd, 1990), mediante el cual el nivel de rasgo no es estimado sino asignado, haciendo que a respuestas correctas le sigan incrementos en el nivel de rasgo y a respuestas erróneas decrementos.

Los métodos bayesianos incorporan información previa sobre la distribución de los niveles de rasgo en la población. De este modo, se añade el condicionante de una cierta distribución a priori de los niveles de rasgo, $g(\theta)$. La distribución a posteriori será:

$$g\left(\theta | u_1 \ldots u_n, x_1 \ldots x_n\right) = \frac{L\left(\theta | u_1 \ldots u_n, x_1 \ldots x_n\right) g(\theta)}{\int L\left(\theta | u_1 \ldots u_n, x_1 \ldots x_n\right) g(\theta) d(\theta)} \quad (3)$$

La estimación Máximo a Posteriori (MAP; Lord, 1986; Mislevy, 1986) busca el nivel de rasgo donde es máxima esta distribución a posteriori:

$$\hat{\theta}_{q-1} = \arg\max_{\theta} g\left(\theta | u_1 \ldots u_n, x_1 \ldots x_n\right). \quad (4)$$

En el caso de una distribución a priori uniforme, MAP y máxima-verosimilitud coinciden.

En la estimación mediante el valor Esperado a Posteriori (EAP; Bock y Mislevy, 1982) el nivel de rasgo estimado es el valor esperado de la distribución a posteriori:

$$\hat{\theta}_{q-1} = \int \theta g\left(\theta | u_1 \ldots u_n, x_1 \ldots x_n\right) d(\theta). \quad (5)$$

Evaluación del criterio de parada

Varios son los criterios de parada que pueden aplicarse en un TAI. De entre éstos, destacan las reglas de: (a) alcanzar un cierto número de ítems administrados; (b) reducir la incertidumbre en la estimación del nivel de rasgo por debajo de un nivel predeterminado, o (c) considerar que la aportación en precisión de ítems adicionales sería inferior a un cierto límite. Combinaciones de estos criterios resultan posibles. Cuando se usa únicamente el criterio (a), estamos hablando de tests de longitud fija; en otro caso, de tests de longitud variable. El criterio de parada a aplicar en un TAI determinado dependerá de diferentes factores. Por ejemplo, cuando multitud de especificaciones respecto al contenido del test han de ser controladas, la opción más razonable es un test de longitud fija, puesto que resulta más sencillo cumplir con los requisitos si se conoce de antemano cuántos ítems van a administrarse. En los casos en los que la validez aparente del test es un elemento importante también suele optarse por longitud fija. En los casos en los que conviene reducir tanto como resulte posible el número de ítems acostumbre a emplearse longitud variable.

Selección de ítems

En este punto puede distinguirse entre dos condiciones, cuando los ítems son seleccionados de uno en uno o cuando los ítems son seleccionados por bloques. En el primer caso, el test tiene tantos puntos de adaptación a la ejecución como ítems van a ser administrados. En el segundo caso, los puntos de adaptación se acostumbran a limitar a tres o cuatro. El término de TAI se suele reservar para el primer caso. El segundo recibe el nombre de Test Multietápico (Luecht y Nungester, 1998). La selección de ítems es uno de los aspectos sobre los que más se ha investigado en el campo de los TAIs. Por ello, se describe con mayor detalle en el siguiente apartado.

Respuesta

Éste es el único elemento del TAI en el que interviene el evaluado. Los ítems que se administran en un TAI han sido calibrados bajo alguno de los modelos de TRI disponibles. Los TAIs permiten incorporar cualquiera de los módelos de TRI propuestos. Al igual que en el campo general de la TRI, la mayor parte de la investigación y aplicaciones se han efectuado sobre modelos dicotómicos. Se ha investigado, también, sobre TAIs para modelos politómicos (Choi y Swartz, 2009; De Ayala, 1992), modelos no paramétricos (Xu y Douglas, 2006), testlets (conjuntos de ítems con un enunciado común y que, por tanto, violan el supuesto de independencia local – Wainer, Bradlow y Wang, 2007), modelos multidimensionales (Luecht, 1996; Mulder y van der Linden, 2009; Segall, 1996) o modelos de diagnóstico cognitivo (Cheng, 2009).

La probabilidad de acierto al ítem viene caracterizada por parámetros propios de los ítems y el nivel de rasgo del examinado. Por eso, los problemas característicos en la TRI de calibración y evaluación del ajuste del modelo (Baker y Kim, 2004) son igualmente pertinentes en los TAIs, con el problemas añadido de que la matriz de respuestas de los examinados a los ítems está casi vacía. Estudios relevantes en este apartado incluirían los relativos a la calibración de nuevos ítems insertados en un TAI operativo (Ban, Hanson, Wang, Yi y Harris, 2001; Ban, Hanson, Yi y Harris, 2002; Chang y Lu, 2010), el estudio del funcionamiento diferencial (Lei, Chen y Yu, 2006, Zwick, 2000) o el estudio de patrones anómalos (Nering, 1997).

Selección de ítems en TAIs

En general, la selección de ítems en un TAI se realiza definiendo un subconjunto del banco de ítems y buscando cuál es el ítem de este sub-banco que optimiza una determina función de valoración. A esta base general se pueden añadir otras restricciones. El modo de determinar ese subconjunto de preguntas y la función de valoración a emplear vendrá determinado por el peso relativo de los diferentes objetivos que ha de satisfacer el test. Por ello, será desarrollando los objetivos de TAI cómo se presentarán las principales reglas de selección de ítems que se han ofrecido.

Davey y Parshall (1995) identifican tres objetivos básicos a cumplir mediante un TAI. A éstos añadiremos un cuarto.

1. Permitir la estimación precisa del nivel de rasgo de los evaluados.
2. Limitar la probabilidad e implicaciones de una filtración de ítems.
3. Garantizar el ajuste a las especificaciones de contenido de la prueba.
4. Facilitar el mantenimiento del banco de ítems.

En los siguientes apartados, desarrollaremos estos puntos más extensamente.

Permitir la estimación precisa del nivel de rasgo de los evaluados

Al igual que con el resto de tests, la interpretación de las puntuaciones de un TAI puede estar orientada a criterio o a norma. En el primer caso, lo que se busca es clasificar a los examinados en una o más categorías (apto o no apto; nivel bajo, medio o alto). Lo relevante no es la precisión en la estimación del nivel de rasgo, sino la precisión y consistencia de las clasificaciones. Dos son las vías básicas para clasificar examinados. La primera, conseguir estimaciones precisas de su nivel de rasgo y comparar éstas con los puntos de corte. La segunda supone centrarse únicamente en si el examinado está por encima o por debajo del punto de corte. Estas dos opciones dan lugar a diferentes modos de seleccionar ítems (Thompson, 2009).

El problema de interpretación referida a norma supone objetivos y, por tanto, criterios diferentes para la selección de ítems. En este caso, buscamos situar a todos los examinados en un continuo con la mayor precisión posible. Nos detendremos en los modos que se han propuesto para determinar qué ítems son los más apropiados para este objetivo.

La mayor parte de las reglas propuestas seleccionan el ítem que optimiza una cierta función de valoración. Esta función puede tomar como entrada un único valor, el nivel de rasgo estimado, o un intervalo de rasgos. La función más comúnmente empleada es la función de información de Fisher evaluada únicamente para el rasgo estimado.

$$I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)\left[1 - P_i(\theta)\right]}, \qquad (6)$$

donde $P_i'(\theta)$ es la primera derivada de $P_i(\theta)$.

El ítem seleccionado es aquel que maximiza la información (Lord, 1980):

$$j = \arg\max_{i \in B_q} I_i\left(\hat{\theta}\right), \qquad (7)$$

donde B_q define aquellos ítems del banco que son evaluados para determinar cuál será administrado en la q-ésima posición. A este criterio lo llamaremos selección por máxima información puntual.

La base de este criterio es una propiedad asintótica de la estimación máximo-verosímil. Bajo ciertas condiciones, el recíproco de la información de Fisher para el nivel estimado equivale al error típico de medida (Chang y Ying, 2009). Por tanto, incrementos en la información reducirán el error.

Esta función de valoración no está exenta de problemas, tanto en términos de precisión como de seguridad. Empezaremos por las limitaciones en la precisión, dejando las de seguridad para una sección posterior. Respecto a la precisión, el inconveniente principal viene causado por seleccionar ítems considerando el nivel estimado como idéntico al nivel real, obviando el error de medida, especialmente presente cuando son pocos los ítems administrados. Un modo de tener en cuenta la incertidumbre sobre la ubicación del nivel de rasgo es ponderar la función de información de Fisher mediante la función de verosimilitud (Veerkamp y Berger, 1997). De este modo, para los primeros ítems, cuando la función de verosimilitud es más plana, se buscarán ítems que ofrezcan información para todo el rango posible de niveles de rasgo, en lugar de información concentrada para un nivel estimado que puede distar en gran medida del rasgo estimado final:

$$j = \arg\max_{i \in B_q} \int I_i(\theta) L\left(\theta | u_1 \ldots u_n, x_1 \ldots x_n\right) d(\theta) \qquad (8)$$

La función de información de Fisher indica la capacidad de un ítem para discriminar entre valores adyacentes de rasgo. Cuando la estimación es pobre, resultaría más oportuna una función que nos permitiera discriminar entre cualesquiera pares de valores. Esto lo permite la función Kullback-Leibler, propuesta como función de valoración en los TAIs por Chang y Ying (1996). No todas las combinaciones de valores de rasgo son igual de probables. Varias han sido las propuestas para incorporar esta incertidumbre en la función. De entre ellas, la más recomendable es ponderar la función

Kullback-Leibler por la función de verosimilitud (Barrada, Olea, Ponsoda y Abad, 2009). Esta función de valoración se expresa del siguiente modo:

$$j = \arg\max_{i \in B_q} \int KL_i(\theta \,||\, \hat{\theta}) L(\theta | u_1 \ldots u_n, x_1 \ldots x_n) d(\theta), \qquad (9)$$

donde

$$KL_i(\theta \,||\, \hat{\theta}) = P_i(\hat{\theta}) \ln\left[\frac{P_i(\hat{\theta})}{P_i(\theta)}\right] + [1 - P_i(\hat{\theta})] \ln\left[\frac{1 - P_i(\hat{\theta})}{1 - P_i(\theta)}\right]. \qquad (10)$$

La solución a las integrales de las Ecuaciones 8 y 10 se aproxima mediante puntos de cuadratura en θ.

Van der Linden (1998) sugiere otra aproximación, las reglas de selección bayesianas. Desde su punto de vista, es oportuno tener en cuenta que una vez administrada una nueva pregunta la estimación del nivel de rasgo cambiará. Estos cambios entre estimaciones sucesivas serán mayores cuanto más próximos al comienzo del test nos encontremos (Chang y Ying, 2008). La idea de van der Linden es buscar el ítem que maximice la información en los niveles de rasgo donde se situaría la estimación en el caso de acierto o error, ponderando por la probabilidad de acierto o error. Para ello, ofrece varias propuestas. A modo de ejemplo, la siguiente:

$$j = \arg\max_{i \in B_q} \left\{ I_i(\hat{\theta}_{u_i=1}) P_i(\hat{\theta}) + I_i(\hat{\theta}_{u_i=0}) [1 - P_i(\hat{\theta})] \right\}, \qquad (11)$$

donde $\hat{\theta}_{u_i=1}$ y $\hat{\theta}_{u_i=0}$ serían los niveles de rasgo estimados en el caso de respuesta correcta o incorrecta al ítem *i*, respectivamente.

Las funciones de valoración alternativas a la de máxima información de Fisher comparten varias características: (a) a medida que aumenta el número de ítems administrados, convergen hacia la seleccion de ítems por máxima información puntual (Ecuacion 7), puesto que la función de verosimilitud es más apuntada y los niveles de rasgo en caso de acierto o fallo están cada vez más próximos entre sí); (b) por tanto, a mayor número de ítems, menor es el beneficio en precisión por el uso de las funciones alternativas en comparación con la selección por máxima información puntual (Chen, Ankenmann y Chang, 2000; Chen y Ankenmann, 2004).

Limitar la probabilidad e implicaciones de una filtración de ítems

Mientras que los examinadores mantienen para todo tipo de test la voluntad de conseguir una estimación precisa del nivel de rasgo de los examinados, el objetivo de los evaluados puede variar según las consecuencias del resultado de la evaluación (Wainer, 2000b). En las situaciones en las que un elevado error de estimación les suponga consecuencias adversas, los examinados desearán un nivel estimado tan próximo a su nivel real como resulte posible (por ejemplo, en ciertas pruebas de diagnóstico psicopatológico donde tanto las falsas alarmas como las omisiones pueden conllevar efectos negativos para los examinados). Pero en otros casos lo más ventajoso para los examinados es conseguir un cierto nivel de rasgo, con independencia de si éste se corresponde o no con el real. Por ejemplo, en una prueba de reválida de Bachillerato en la que no superarla supone la no homologación de varios años de estudio. En este caso, la mayor parte de los evaluados deseará un resultado de apto con independencia de su nivel real.

En los procesos de evaluación en los que los objetivos de examinadores y examinados no concuerdan es probable que una parte de los evaluados busquen el modo de incrementar artificialmente sus calificaciones. En una revisión de prácticas tramposas en exámenes, Cizek (1999) señaló que la mitad o más de los estudiantes admiten copiar. Varias son las opciones para esto (Davey y Nering, 2002). Por ejemplo, mirar alrededor buscando algún examinado con una pregunta idéntica a la nuestra y copiarle la respuesta o llevar auriculares y micrófonos ocultos para que alguien vaya enviando desde el exterior las respuestas. Una alternativa especialmente provechosa es llevar conocido de antemano parte del banco de ítems, de tal modo que la probabilidad de respuesta correcta a esas preguntas sea elevada, con independencia del parámetro de rasgo. Como señal de lo útil de esta estrategia, un 76% de los estudiantes universitarios preguntados por Stern y Havlick (1986) reconocieron preguntar a estudiantes ya evaluados por el contenido de su examen.

El riesgo de conocimiento previo de ítems está especialmente presente en los TAIs (Chang, 2004). Esto se debe a dos motivos básicos:

- Los TAIs cobran sentido cuando se ofrecen de forma continuada. A diferencia de algunos programas en los que únicamente se administran las pruebas unas pocas ocasiones al año, la mayor parte de los TAIs permanecen activos todo el año. En las pruebas que no se ofrecen de un modo continuo, los ítems empleados son, por lo general, automáticamente descartados para ocasiones posteriores. En los TAIs, los bancos de ítems son relativamente estáticos a lo largo del tiempo. Esto supone que un examinado puede preguntarle a personas previamente evaluadas por los ítems que recibieron. También puede recurrir a academias de preparación de exámenes o sitios web especializados en el filtrado de preguntas.
- Conocer ciertos ítems de antemano será provechoso si algunos de estos le son presentados al examinado. La mayor parte de las reglas de selección implican una alta probabilidad de que haya coincidencia entre los ítems preconocidos y aquellos administrados. Por un lado, al comienzo del test los niveles de rasgo estimados tienden a ser muy parecidos para los diferentes examinados. En general, a mismo nivel de rasgo estimado, mismo ítem administrado. Por otro lado, puesto que las reglas de selección de ítems buscan aquellos ítems con mejores propiedades métricas, los ítems seleccionados se concentran entre aquellos de mayor parámetro de discriminación (Barrada, Olea et al., 2009; Li y Schafer, 2005).

La consecuencia de esto es una elevada coincidencia entre los ítems recibidos por diferentes examinados. Esto conlleva que contar con información previa al test sobre parte

del contenido del banco de ítems puede servir para inflar el nivel estimado de los examinados (Yi, Zhang y Chang, 2008). Este riesgo puede llegar a tener consecuencias aplicadas, como la suspensión de importantes programas de evaluación mediante TAIs (Chang, 2004; Honan, 1995).

La tasa de solapamiento (la proporción de ítems que dos examinados tomados al azar comparten – Way, 1998) y la distribución de las tasas de exposición se han convertido en las dos variables más habituales para evaluar el riesgo de conocimiento previo (Chang y Zhang, 2002). Se ha entendido que a mayor tasa de solapamiento, más provechoso puede ser el preconocimiento de ítems. Igualmente, se ha considerado que distribuciones más homogéneas de tasas de exposición son preferibles. Chen, Ankenmann y Spray (2003) han mostrado que la tasa de solapamiento, para tests de igual longitud e igual tamaño del banco de ítems, crece linealmente con la varianza de las tasas de exposición de los ítems. La tasa de solapamiento puede calcularse mediante la siguiente ecuación:

$$T = \frac{n}{Q} S^2_{P(A)} + \frac{Q}{n}, \qquad (12)$$

donde n es el tamaño del banco, Q es la longitud del test y $S^2_{P(A)}$ es la varianza de las tasas de exposición.

Varias son las propuestas que se han formulado hasta el momento para limitar este riesgo.

Restricciones en el banco presentable

Entendemos por banco presentable (B_q) el subconjunto de preguntas del banco de ítems que son evaluadas por la regla de valoración a la hora de seleccionar el ítem para su administración en la posición q. La restricción mínima a la hora de configurar B_q es no incluir aquellos ítems administrados al examinado en posiciones previas del test. Una parte importante de las propuestas para mejorar la seguridad del banco parten de limitar el tamaño de B_q, de tal modo que se reduzca la probabilidad de que aquellos ítems con tendencia a ser más expuestos formen parte de él. Cuatro son los modos básicos de hacer esto:

- *Restricción de tasa máxima de exposición:*

Una opción para homogeneizar las tasas de exposición de los ítems es limitar la proporción de examinados que pueden recibir los ítems. A esta tasa máxima, fijada de antemano por la institución o empresa responsable del test, la llamaremos r^{max}. El modo de limitar la tasa máxima es reducir la frecuencia con la que los ítems con tasas de exposición elevadas entran en B_q (Barrada, Abad y Veldkamp, 2009).

El primer método propuesto para esto es el método Sympson-Hetter (Hetter y Sympson, 1997; Sympson y Hetter, 1985). En la descripción del mismo seguiremos la reformulación de Barrada, Abad y Veldkamp (2009), equivalente en resultados a la original, si bien computacionalmente más rápida. Esta formulación es más próxima a propuestas posteriores de control de r^{max}, lo cual facilita la comparación entre ellas. Sean dos eventos diferentes: (a) E_i, el ítem i es marcado como elegible (incorporado a B_q); y (b) A_i, el ítem i es administrado. Puesto que cualquier ítem administrado ha de ser elegible, se cumple que:

$$P(A_i) = P(A_i \mid E_i)P(E_i). \qquad (13)$$

$P(A_i)$, la probabilidad de administración o tasa de exposición, es el valor que quiere controlarse. El modo de conseguirlo es fijando convenientemente valores de $P(E_i)$, la probabilidad de que un ítem entre en B_q. Estas probabilidades de elegibilidad se calculan mediante un proceso iterativo. Los parámetros $P(E_i)$ para el ciclo $t+1$ derivan de hacer $P(A_i)$ igual a r^{max} en la ecuacion anterior y fijar en 1 el valor máximo de estos parámetros:

$$P^{(t+1)}(E_i) = \begin{cases} 1 & si \quad P^{(t)}(A_i)/P^{(t)}(E_i) \le r^{\max} \\ r^{\max} P^{(t)}(E_i)/P^{(t)}(A_i) & si \quad P^{(t)}(A_i)/P^{(t)}(E_i) > r^{\max} \end{cases} \qquad (14)$$

Cuando el valor de la tasa máxima de exposición se estabiliza o cuando se alcanza un número predefinido de ciclos, el proceso de simulación para establecer los parámetros $P(E_i)$ finaliza. Una vez establecidos estos valores, para cada examinado se generan tantos números aleatorios dentro del intervalo uniforme (0, 1) como ítems componen el banco. Sólo en el caso de que el número aleatorio sea menor a $P(E_i)$ el ítem i formará parte de B_q.

Este método presenta varios problemas. Primero, no todas las tasas quedan por debajo del límite (van der Linden, 2003). Segundo, el método funcionará en la medida en la que la distribución de rasgo de la simulaciones coincidan con la distribución de rasgo de los examinados (Chen y Doong, 2008). Tercero, las simulaciones han de ser repetidas cada vez que se incorpora o se retira cualquier ítem del banco (Chang y Harris, 2002). Cuarto, las simulaciones necesarias consumen tiempo, si bien se han propuesto varias vías para reducirlo (Barrada, Olea y Ponsoda, 2007; Chen y Doong, 2008; van der Linden, 2003).

Otra alternativa es el método restringido de Revuelta y Ponsoda (1998). Este método, a diferencia del método Sympson-Hetter, no requiere simulaciones previas, sino que ajusta la pertenencia o no de cada ítem a B_q para cada nuevo examinado, según las tasas de exposición encontradas. Aquellos ítems con una tasa de exposición mayores de r^{max} hasta un cierto examinado son retirados de B_q y no vuelven a ser incorporados hasta que su tasa se sitúa por debajo del límite. Con este método, la pertenencia o no a B_q es determinista, no probabilística como en el método Sympson-Hetter, y se ajusta para cada nuevo examinado. Siendo f el indicador de la posición ordinal del examinado dentro del total de examinados evaluados, los valores de $P(E_i)$ se ajustan mediante la siguiente fórmula:

$$P^{(f+1)}(E_i) = \begin{cases} 1 & si \quad P^{(1..f)}(A_i) < r^{\max} \\ 0 & si \quad P^{(1..f)}(A_i) \geq r^{\max} \end{cases} \qquad (15)$$

En el método de elegibilidad del ítem (van der Linden y Veldkamp, 2004) la pertenencia a B_q no es determinista. Este sistema, al igual que el método restringido, no requiere de simulaciones previas y la probabilidad de entrar en B_q no es fija para todos los examinados, sino que se va adaptando según la administración o no del ítem a examinados previos:

$$P^{(f+1)}(E_i) = \\ = \begin{cases} 1 & si \quad P^{(1..f)}(A_i) / P^{(f)}(E_i) \leq r^{\max} \\ r^{\max} P^{(f)}(E_i) / P^{(1..f)}(A_i) & si \quad P^{(1..f)}(A_i) / P^{(f)}(E_i) > r^{\max} \end{cases} \qquad (16)$$

De entre las tres propuestas descritas, la alternativa ofrecida por van der Linden y Veldkamp (2004) es la que parece ser preferible (Barrada, Abad y Veldkamp, 2009): (a) no necesita simulaciones previas para calcular los parámetros de control de la exposición; (b) es independiente de supuestos sobre la distribución del nivel de rasgo de los examinados; y (c) satisface de un modo casi completo la restricción de que no haya ninguna tasa de exposición mayor a r^{max}.

Cuando se combina la regla de selección de ítems más comúnmente empleada, la de máxima información puntual, con los métodos de control de r^{max}, los ítems de mayor parámetro de discriminación son administrados al comienzo del test. Suponiendo que todos los examinados comenzaran el TAI con el mismo nivel de rasgo estimado, el ítem más informativo para ese nivel de rasgo sería presentado el "r^{max} por ciento" de las veces como primer ítem del test; igualmente, ese ítem nunca sería presentado en una posición que no fuera la primera. Sin embargo, diferentes estudios han mostrado que una selección de ítems altamente basada en el azar al comienzo del test apenas deteriora la estimación final de los niveles de rasgo (Li y Schafer, 2005; Revuelta y Ponsoda, 1998). Teniendo esto en cuenta, Barrada, Veldkamp y Olea (2009) han propuesto el método de múltiples tasas máximas. En este método, que toma como base el modo de calcular los parámetros de control de la exposición de los ítems del método de elegibilidad del ítem (Ecuación 16), se determinan tantos valores de r^{max} como ítems van a ser administrados. La tasa máxima de exposición de los ítems al comienzo del test es tan baja como resulte posible dado el tamaño del banco de ítems. El valor de r^{max} condicionado a la posición del ítem en el test se incrementa según avanza el test. De este modo, resulta posible mejorar la seguridad del banco (reducciones en la tasa de solapamiento) con incrementos despreciables en el error de medida.

El control de la tasa máxima tal y como se efectúa mediante los diferentes métodos descritos deja la opción de sobreexposición de ciertos ítems y alto solapamiento cuando se condicionada a niveles de rasgo (Davey y Parshall, 1995). Por ejemplo, sería posible que casi todos los examinados con alto nivel de rasgo, infrecuentes en la población, recibieran idénticos ítems y, aún así, la tasa de exposición de éstos estuviera por debajo de r^{max}. Por eso, se ha propuesto controlar la tasa máxima de exposición condicionada a nivel de rasgo, tanto real (Stocking y Lewis, 1998, para una variante del método Sympson-Hetter) como estimado (Stocking y Lewis, 2000, para el mismo método; van der Linden y Veldkamp, 2007, para el método de elegibilidad del ítem). En este caso, a cada ítem ya no le corresponde un único parámetro de control de la exposición, sino tantos parámetros como intervalos en los que hayamos dividido el continuo del nivel de rasgo.

- *Restricciones de tasa máxima de exposición y tasa de solapamiento:*

Distintas funciones de valoración de ítems, con la misma restricción de la tasa máxima, no conllevan la misma tasa de solapamiento. No existe una función que determine para cierto nivel de r^{max} qué tasa de solapamiento se obtendrá. Chen y Lei (2005) ampliaron el método Sympson-Hetter para permitir el control simultáneo de r^{max} y la tasa de solapamiento. Este método se basa en la relación conocida entre varianza de las tasas de exposición y solapamiento (Ecuación 12). Fijando la tasa de solapamiento objetivo (T^{obj}), es posible calcular la varianza de las tasas de exposición que llevaría a T^{obj}. A esta varianza la llamaremos S^{obj}. En cada nueva iteración del ciclo de simulaciones requerido para calcular los parámetros de control de la exposición, la tasa de exposición objetivo para los ítems en el ciclo *t+1* se sitúa en:

$$P^{(t+1)}(A_i) = S^{obj}\left(\frac{P^{(t)}(A_i) - \frac{Q}{n}}{S_{(t)}}\right) + \frac{Q}{n}, \qquad (17)$$

donde $S_{(t)}$ es la desviación típica de las tasas de exposición para el ciclo *t*. De este modo, se fija la distancia en desviaciones típicas con respecto a la tasa media (Q/n) para los diferentes ítems de ciclo en ciclo. El proceso de ciclos de simulación permite ir afinando los parámetros de control de la exposicíon hasta conseguir el control de r^{max} y T^{obj}.

Desarrollos posteriores de esta idea han permitido que este control simultáneo se pueda hacer sin simulaciones previas para conseguir las probabilidades de control de la exposición (Chen, Lei y Liao, 2008). La última aportación permite controlar la tasa de solapamiento entre grupos de examinados (Chen, 2010), no únicamente entre pares de examinados.

- *Bancos rotatorios:*

Algunos programas de evaluación cuentan con bancos de miles de ítems. Esto permite construir diferentes sub-bancos de menor tamaño, todos ellos con de la misma capacidad de medida para los diferentes niveles de rasgo (Mills y Steffen, 2000). Hay varias estrategias para manejar los sub-bancos: (a) pueden irse alternando por emplazamientos de evaluación, en aquellos programas en los que la evaluación se realiza desde un número limitado de centros; (b) pueden irse activando y desactivando según una programación temporal; o (c) al comenzar el test, puede asignarse aleatoriamente a cada examinado el banco que le corresponderá. De este modo, se limita la tasa máxima de exposición de los ítems (un ítem no puede tener mayor tasa que la proporción de ocasiones en

las que se emplea el sub-banco o sub-bancos a los que pertenece).

Dos son los diseños de bancos rotatorios, con coincidencia de ítems o sin ella (Ariel, Veldkamp y van der Linden, 2004). Los bancos sin coincidencia son aquellos en los que la extracción de ítems del banco maestro (aquel que contiene todos los ítems) se realiza sin reposición, de tal modo que los sub-bancos no comparten ítem alguno. Los bancos con coincidencias permiten que aquellos ítems con menor tasa de exposición formen parte de varios sub-bancos.

Cuando se han comparado la estrategia de bancos rotatorios y de restricción de r^{max}, se ha encontrado que sus resultados en precisión y seguridad son casi indistinguibles (Barrada, Olea y Abad, 2008), con el método de los bancos rotatorios superando ligeramente a la opción del control de tasas máximas.

- *Métodos estratificados:*

Los métodos de restricción de r^{max} (con o sin control de la tasa de solapamiento) afrontan el problema de la seguridad del banco reduciendo la sobreexposición de los ítems más populares. Restricciones de r^{max} incrementan la exposición de aquellos ítems ligeramente por debajo en calidad métrica de aquellos cuya exposición se limita. Este modo de actuar tiene un efecto más bien limitado a la hora de incrementar la tasa de exposición de los ítems nunca o apenas utilizados.

Los métodos estratificados son una propuesta para incrementar el uso de aquellos ítems infraexpuestos (Chang y Ying, 1999). En éstos, B_q se hace variable según la posición del ítem dentro de la secuencia de preguntas en el test. Al comienzo del test, cuando el error de medida es máximo y la estimación del nivel de rasgo es más inestable, B_q está compuesto únicamente por aquellos ítems con menor capacidad discriminativa. A medida que el test avanza, B_q se va componiendo por ítems de mayor calidad. De este modo: (a) se fuerza que ítems que no serían nunca empleados lo sean, ya que son los únicos disponibles; y (b) se reservan los ítems de mayor discriminación para las fases finales del test, cuando la estimación es más precisa.

Cambios en la función de valoración de ítems

Las funciones de valoración que hemos revisado buscan maximizar la precisión de medida. Otras funciones, sin embargo, han sido desarrolladas con la idea de incrementar la seguridad del banco de ítems. La idea básica para ello es reducir la sobreexposición de los ítems con alto parámetro de discriminación e incrementar las tasas de exposición de aquellas preguntas que, con las reglas orientadas a la precisión, nunca o casi nunca son presentadas. De las funciones de valoración propuestas con este fin, destacamos:

- *Los métodos de distancia con respecto a la dificultad del ítem:*

Una opción de romper la relación entre parámetro de discriminación y tasa de exposición es seleccionar los ítems atendiendo únicamente a su parámetro de localización, sin tener en cuenta su capacidad discriminativa. Un ítem calibrado según un modelo dicotómico ofrece información máxima para valores de rasgo iguales al parámetro de localización (modelo de 1 y 2 parámetros) o ligeramente por encima de éste (modelo de 3 parámetros). En el modelo de 3 parámetros, el nivel de rasgo en el que un ítem alcanza su información de Fisher máxima se sitúa en (Hambleton y Swaminathan, 1985):

$$\theta_i^{\max} = b_i + \frac{\ln\left[1+\left(1+8c_i\right)^{1/2}\right]-\ln(2)}{1.7a_i}. \quad (18)$$

Una posible función de valoración sería, pues, seleccionar los ítems con distancia mínima entre el nivel de rasgo estimado y su parámetro *b*:

$$j = \arg\min_{i\in B_q}\left|\hat{\theta} - b_i\right|. \quad (19)$$

O entre el nivel de rasgo estimado y el nivel de rasgo en el que se obtiene la máxima información del ítem:

$$j = \arg\min_{i\in B_q}\left|\hat{\theta} - \theta_i^{\max}\right|. \quad (20)$$

De este modo, estaremos: (a) equilibrando en gran medida las tasas de exposición de los ítems (Li y Schafer, 2005); y (b) administrado los ítems a aquellos examinados para los que resultan más convenientes. El problema es una reducción en la precisión de medida, al considerar equivalentes en la selección ítems de diferente capacidad discriminativa.

Por eso, los métodos de distancia con respecto a la dificultad suelen a aplicarse combinados con la estrategia de banco presentable variable según posición del ítem en el test. El método alfa-estratificado (Chang y Ying, 1999) se ajusta a este perfil y ha sido, probablemente, la propuesta que ha recibido más atención en los últimos años en la investigación de la seguridad en TAIs. En este método, la capacidad de discriminación de los ítems aumenta según avanza el test y la selección se realiza teniendo en cuenta únicamente el rasgo estimado y el parámetro *b*. Barrada, Mazuela y Olea (2006) propusieron el equivalente al método alfa-estratificado cuando se incluye el cambio en la ubicación del nivel de máxima información que introduce el parámetro de pseudo-azar en el modelo de 3 parámetros. Igualmente, propusieron estratificar el banco no según el parámetro *a*, sino según la información máxima que puede alcanzar un ítem (Hambleton y Swaminathan, 1985):

$$I_i^{\max} = \frac{1.7^2 a_i^2}{8\left(1-c_i^2\right)}\left[1-20c_i-8c_i^2+\left(1+8c_i\right)^{3/2}\right]. \quad (21)$$

- *El método progresivo:*

Esta propuesta, de Revuelta y Ponsoda (1998), supone empezar el test con selección aleatoria de ítems y, según la

prueba va avanzando, ir incrementando el peso de la información en la selección de ítems. Este modo de proceder permite reducir la infraexposición (ningún ítem tiene una tasa de exposición nula, ya que el test comienza con selección al azar) y también reducir ligeramente la sobreexposición (parte del problema de la sobreexposición proviene del limitado número de posibles niveles de rasgo estimados al comienzo del test, lo que dispara la demanda de ítems para estos pocos niveles). Esto se consigue con una perdida nula o muy pequeña en la precisión de medida.

Concretamente, en el método progresivo la función se valoración se compone de dos elementos, uno aleatorio (R_i) dentro del intervalo $[0, max_{i\in B_q} I_i(\hat{\theta})]$ y el otro, la información de Fisher del ítem:

$$j = \arg\max_{i\in B_q} \left|\left(1-W_q\right)\mathbb{R}_i + W_q I_i\left(\hat{\theta}\right)\right|. \qquad (22)$$

El peso del componente de la información (W_q) varía según la posición del ítem en el test. Revuelta y Ponsoda (1998) propusieron para definir W_q:

$$W_q = \frac{q-1}{Q}, \qquad (23)$$

donde q es la posición serial del ítem y Q la longitud del test. Esta ecuación supone una reducción lineal del componente aleatorio, empezando por selección aleatoria y acabando por selección casi únicamente determinada por la información de Fisher. Otra opción, planteada por Barrada, Olea, Ponsoda y Abad (2008), flexibiliza el modo de definir la relación entre el componente aleatorio y la posición del ítem en el test. Esto se realiza mediante la siguiente fórmula:

$$W_q = \begin{cases} 0 & si \quad q=1 \\ \dfrac{\sum_{h=1}^{q}(h-1)^t}{\sum_{h=1}^{Q}(h-1)^t} & si \quad q\neq 1 \end{cases}. \qquad (24)$$

En este ecuación, el parámetro t permite establecer la velocidad con la que se reduce el componente aleatorio. Un valor igual a 0 supone una transición lineal desde W_q igual a 0 hasta W_q igual a 1. Incrementos en el valor de t conllevan aumentar la presencia de aleatoriedad en la selección de ítems. De este modo, se puede buscar el parámetro que permite aumentar al máximo la seguridad sin que ello implica pérdida en la precisión de medida.

Varias son las razones que explican cómo es posible administrar ítems al azar y que esto no incremente el error de medida:

- Los ítems administrados por el método de máxima información puntual van siendo cada vez menos informativos, mientras que con el método progresivo el patrón es el inverso, haciendo que la información acumulada al final del test sea equivalente.
- El método de máxima información puntual presenta problemas cuando, al comienzo del test, examinados de alto nivel de rasgo fallan un ítem o examinados de rasgo bajo lo aciertan (Rulison y Loken, 2009). En tales casos, el nivel estimado se desplaza hacia el extremo incorrecto y, puesto que los ítems administrados son de alto parámetro *a*, la función de verosimilitud es muy apuntada. Así las cosas, es necesaria la presentación de una gran cantidad de ítems hasta conseguir que el nivel estimado se aproxime al nivel real. El método progresivo, dada la importancia de la selección aleatoria, reduce el impacto de estos patrones anómalos al comienzo del test.

- *Métodos estocásticos:*

Tal y como apuntábamos anteriormente, lo habitual es que la selección de ítems se efectúe determinando, entre los ítems que forman B_q, aquel que optimiza una determinada función de valoración. Conociendo el nivel de rasgo estimado del examinado y la composición de B_q, podemos determinar con certeza cuál será el siguiente ítem a administrar. Esto supone que aquellos ítems que nunca se encuentren entre los Q primeros ítems según esa función de optimización nunca serán presentados, independientemente de sus propiedades psicométricas. Una alternativa sería emplear procedimientos estocásticos de selección de ítems (Segall, 2004). Por ejemplo, Barrada, Olea, Ponsoda y Abad (2008) proponen el método proporcional, en el que la información de Fisher para el nivel de rasgo estimado elevada a una cierta potencia sirve para calcular la probabilidad de selección. Al comienzo del test, la selección de ítems es aleatoria. Para el último ítem a administrar, la selección mediante el método proporcional o el método de máxima información puntual será muy similar. La potencia a la que se eleva la información de Fisher (P_q) aumenta para cada nuevo ítem administrado, siguiendo la siguiente ecuación:

$$P_q = \begin{cases} 0 & si \quad q=1 \\ \dfrac{Q\sum_{h=1}^{q}(h-1)^s}{\sum_{h=1}^{Q}(h-1)^s} & si \quad q\neq 1 \end{cases} \qquad (25)$$

El parámetro s desempeña un papel similar el del parámetro t en el método progresivo (Ecuación 24). Valores elevados de s suponen mantener durante una parte importante del test un alto componente del azar en la selección. De este modo, se puede incrementar importantemente la seguridad del banco sin efectos apreciables en la precisión.

- *Métodos mixtos:*

No hay razón alguna por la que haya que mantener constante la función de valoración de ítems a lo largo de todo el test. Distintas propuestas se han planteado de combinaciones de funciones, todas ellas compartiendo la idea de empezar por métodos menos precisos pero más seguros y, según el test avanza, pasar a funciones donde la información desempeña un papel más importante. Li y Schafer (2005) proponen empezar por selección aleatoria, pasar a selección

según distancia a dificultad y acabar por máxima información. Leung, Chang y Hau (2005) y Barrada, Abad y Olea (2011) optan por estratificar el banco, empezar por selección según distancia a dificultad y acabar, también, por máxima información puntual.

Como se puede apreciar, hay una amplia variedad de métodos disponibles para la mejora de la seguridad del banco en TAIs (Georgiadou, Triantafillou y Economides, 2007). El patrón de resultados habitualmente encontrado señala la existencia de un balance entre precisión y seguridad, de tal modo que incrementos en un objetivo supone decrementos en el otro (Chang y Ansley, 2003; Finkelman, Nering y Roussos, 2009; Stocking y Lewis, 2000), por lo que resulta complicado determinar qué método es, globalmente, más adecuado (Barrada, Olea, Ponsoda y Abad, 2010).

Restricciones de contenido de los tests

Ítems pertenecientes a una misma dimensión pueden variar en los subdominios que cubren (p. ej., en un test de matemáticas puede haber ítems de aritmética, de probabilidad y de trigonometría). Previamente a la puesta en funcionamiento del programa de evaluación, la agencia encargada del test desarrolla una tabla de especificaciones en la que se detalla el rango de ítems que cada examinado ha de responder por subdominio.

La agencia responsable del test puede desear controlar otra multitud de aspectos de los ítems. Por ejemplo, impedir la presencia simultánea para un examinado en su test de 'ítems enemigos', aquellos que, de presentarse uno, no ha de administrarse el otro. En algunos casos se controla la proporción de ocasiones en las que la respuesta correcta corresponde a cada una de las alternativas de respuesta. La importancia dentro del test de enunciados referidos a distintos sexos o diferentes etnias también puede ser un elemento a controlar. Algunos programas de evaluación buscan igualar los tests de los diferentes examinados en número de palabras o en tiempo de evaluación. Como se ve, la amplitud y variedad de las restricciones a imponer en un TAI puede ser muy amplia. De hecho, la dificultad para integrar en un TAI todas las restricciones de contenido deseadas ha sido uno de los argumentos empleados para justificar los Tests Multietápicos (Hendrickson, 2007). Por ello, ésta ha sido un área de desarrollo especialmente relevante en el campo de los TAIs. Varias han sido las propuestas para conseguir cumplir con las restricciones formuladas. Nos centraremos en las cuatro principales.

Métodos de espiralización

Estos métodos, inicialmente propuestos por Kingsbury y Zara (1991), sirven únicamente cuando la restricción es respecto a un único atributo categórico y cuando la especificaciones definen un único valor como admisible, no un rango de valores. En la propuesta original, para cada categoría y previamente a la selección de cada ítem, se calcula la división del número de ítems administrados entre el número de ítems a administrar por categoría. El ítem a presentar será seleccionado entre los pertenecientes a la categoría con menor resultado en la división, esto es, máxima discrepancia entre el estado actual y el objetivo. Una variante es utilizar el resultado de esas divisiones para construir una distribución multinomial y asignar aleatoriamente la categoría de la que se escogerá el ítem (Chen y Ankenmann, 2004). El inconveniente básico de este método es lo limitado de los casos en los que se puede aplicar. Su mayor ventaja, la simplicidad.

Modelo de desviación ponderada

En esta propuesta de Stocking y Swanson (1993) cada uno de los atributos a controlar recibe una ponderación determinada por expertos. Los límites especificados en el diseño del test dejan de ser considerados como objetivos estrictos y pasan a ser objetivos deseables. La precisión en la medida de los examinados es considerada también un objetivo con su correspondiente peso. Los ítems seleccionados son aquellos que minimizan la desviación entre los objetivos del test y lo obtenido en el caso de administrar tal ítem:

$$j = \arg\min_{i \in B_q} \sum_{h=1}^{H} z_h \left| \pi_{i,h} - \gamma_h \right|, \qquad (26)$$

donde H es el número de restricciones incluidas en el TAI, z_h es el peso asignado a cada una de estas restricciones, $\pi_{i,h}$ es el valor para la restricción h en el caso de ser administrado el ítem i y γ_h es el objetivo para la restricción h.

Este procedimiento requiere que se determinen los pesos por objetivo (con un cierto componente de ensayo y error) y no garantiza que se satisfagan las especificaciones por completo.

Aproximación del test en la sombra

La aproximación del test en la sombra (van der Linden, 2000; van der Linden y Reese, 1998), basada en métodos de programación lineal, construye para cada nuevo ítem a seleccionar un test completo tal que: (a) se satisfagan todas las restricciones; (b) contenga todos aquellos ítems ya administrados; y (c) sea el test óptimo según la regla de selección. El ítem administrado será aquel que, formando parte del test en la sombra y no habiendo sido administrado al examinado, resulte óptimo según la regla de selección. El test resultante cumple por completo las restricciones y es el más adecuado desde el punto de vista de la regla de selección. La principal diferencia con respecto a otros métodos de restricción de contenidos es que, mientras que en éstos los ítems se seleccionan de uno en uno, el método del test en la sombra construye, para cada nuevo ítem a administrar, un test completo. Esto garantiza que el test administrado sea óptimo.

Método del índice de máxima prioridad

Ésta es la propuesta más reciente (Cheng y Chang, 2009; Cheng, Chang, Douglas y Guo, 2009). Antes de la selección de un ítem, se calcula la 'cuota restante' ($\tau_{i,h}$) para cada restricción, que es la proporción de ítems que falta por administrar correspondientes a esa especificación. Los valores de $\tau_{i,h}$ para aquellas restricciones que el ítem no puede cubrir son fijados a 1. Para cada uno de los ítems que componen el banco, se calcula el producto de las cuotas restantes y este valor, a su vez, se multiplica por la función de valoración del ítem según la regla de selección (V_i). La pregunta con resultado máximo tras estas operaciones es la administrada, dado que ofrece simultáneamente mejor resultado combinado en la función de valoración y en la satisfacción de las restricciones:

$$j = \arg\max_{i \in B_q} V_i \prod_{h=1}^{H} \tau_{i,h} \,. \qquad (27)$$

Al tratarse de un producto, una vez que un objetivo ha sido satisfecho ningún ítem adicional que cubra tal requisito puede ser administrado. Por esto, el método del índice de máxima prioridad supone una satisfacción plena de los restricciones del test.

El método de espiralización es sencillo, pero limitado en su aplicabilidad. El modelo de desviación ponderada ha sido durante años el método preferido por las empresas y agencias responsables de los TAIs, si bien la selección secuencial de ítems no es la más eficaz y, como hemos dicho, puede llevar a que no se cumplan las especificaciones. La aproximación del test en la sombra es versátil, la selección simultánea de ítems es superior a la secuencial y el método es capaz de incorporar tantas restricciones como resulten necesarias sin necesidad de asignarles pesos. Cuando se ha comparado la aproximación del test en la sombra con el modelo de desviación ponderada se ha encontrado que ambos presentan resultados equivalentes en precisión de medida, si bien la primera opción ofrece un mejor control de contenidos (van der Linden, 2005). El problema del test en la sombra es su mayor complejidad, tanto matemática como de programación. El método del índice de máxima prioridad, que empieza a estudiarse recientemente, es eficaz y sencillo para la aplicación de las restricciones, si bien supone una pérdida en la precisión de medida. Al comparar el método del índice de máxima prioridad con el modelo de desviación ponderada, Cheng y Chang (2009) encontraron que ambos métodos eran equivalentes en precisión, si bien el segundo se ajustaba mejor a las especificaciones del test. Ahora bien, el estudio de Cheng y Chang hay que tomarlo con cautela, puesto que para el método de desviación ponderada no incluyeron restricción de tasa máxima de exposición, mientras que sí que lo hicieron para el método del índice de máxima prioridad, por lo que la comparación se realiza sobre métodos que difieren en un aspecto clave en el diseño del TAI. Por el momento, la propuesta de van der Linden y la de Cheng no han sido comparadas. Es probable que su funcionamiento relativo dependa de la cantidad y complejidad de las restricciones a incorporar. Mientras que la aproximación del test en la sombra ha demostrado poder satisfacer 400 especificaciones (van der Linden, 2005), el método del índice de máxima prioridad ha sido puesto a prueba con únicamente 21 restricciones (Cheng y Chang, 2009).

Facilitar el mantenimiento del banco de ítems

Todo banco de ítems requiere de un cierto mantenimiento (Mills y Stocking, 1996). Con el tiempo, el contenido de los constructos puede variar, haciéndose necesario el diseño de preguntas nuevas y la supresión de algunas antiguas. En algunos países existe legislación que obliga a hacer accesible al público parte del contenido del banco, para que futuros examinados reduzcan su incertidumbre sobre el contenido de la prueba. Al así hacerlo, estos ítems pasan a ser inservibles y han de desarrollarse y calibrarse otros que los reemplacen. Aspectos relativos a la seguridad también invitan a la retirada de ítems. Aquellos que han sido utilizados más allá de un cierto límite (sea este límite un tiempo de pertenencia al banco o un número de examinados que han recibido la pregunta) son suprimidos y, si no deseamos reducir el tamaño del banco, han de ser reemplazados por otros. El coste de cada nuevo ítem dependerá de multitud de factores. Según Buyske (2005), puede fácilmente superar los 100 dólares por pregunta. Luecht (2005) lo sitúa más allá, desde varios cientos hasta más de mil quinientos dólares por ítem.

Las funciones de valoración de ítems comúnmente empleadas en los TAIs tienden a dificultar el mantenimiento del banco. Estas funciones de valoración suelen implicar un uso intensivo de aquellos ítems de mayor parámetro de discriminación. Si por cada ítem retirado (situado en la cola derecha en la distribución de discriminación) introducimos un nuevo ítem en el banco (cuya discriminación esperada será igual a la discriminación promedio del banco), lo que estaremos haciendo será ir reduciendo progresivamente la capacidad de discriminación de nuestro banco (Hau y Chang, 2001). Para evitar este deterioro, la ratio entre el número de ítems a calibrar para substituir a los eliminados y el número de ítems descartados tendrá que ser marcadamente mayor de 1. Por otro lado, una parte importante de las reglas de selección de ítems empleadas en los TAIs conlleva que gran parte del banco de ítems sea trivial, en el sentido de que nunca es administrado a examinado alguno. Podemos tener, pues, un gran coste para mantener la calidad del banco y un retorno nulo o mínimo de la inversión para el desarrollo de una parte importante del mismo. Esta situación puede disparar el coste de mantenimiento del programa de evaluación (Wainer, 2000b).

Importancia relativa de los diferentes objetivos

La relevancia de los cuatro objetivos descritos dependerá de multitud de factores específicos de cada programa de eva-

luación. Así, por ejemplo, la fiabilidad deseada de las puntuaciones depende del uso que se le quiera dar a las mismas, al igual que la seguridad del banco de ítems es un asunto menor en aquellos tests en los que los examinados no tienen motivos para falsear sus respuestas.

En líneas generales, el primer objetivo, fiabilidad, y el segundo y el cuarto, seguridad y mantenimiento, son contrapuestos: incrementos en la satisfacción de uno suponen decrementos en los resultados en los otros. Una gran proporción de ítems de un banco no son óptimos para ningún nivel de rasgo en términos de información aportada. Por ello, cuando se prioriza la fiabilidad como criterio, una parte importante del banco no es usada para ningún examinado. Este modo de proceder entra en conflicto con los objetivos de seguridad y mantenimiento. El control de contenidos tendrá un efecto menor en el resto de objetivos, asumiendo un banco de ítems bien construido en el que la composición del banco refleje las restricciones de contenido que se van a imponer.

Conclusiones

Hace años, la transición de un test convencial a un TAI suponía un cambio doble: primero, informatizar el test, y, segundo, hacerlo adaptativo. A medida que los tests informatizados se van popularizando, también lo van haciendo los TAIs. Por tanto, creemos que los relativamente pocos TAIs operativos en nuestro país son los primeros de una larga lista que ha de venir. No en vano, este modo de administración de pruebas presenta claras ventajas: rapidez y precisión.

El concepto básico de un TAI, tanto en términos psicométricos como de programación informática, es relativamente sencillo. La complejidad entra a medida que objetivos adicionales a la precisión de medida van siendo incorporados. Ha sido nuestra intención con este artículo el facilitar una revisión actualizada y en castellano sobre el campo, haciendo más accesible los retos y soluciones propuestas cuando se opta por un tests adaptativo. Como toda revisión, es parcial, puesto que el pequeño mapa que ofrecemos ha de ser más pequeño que el territorio completo de los TAIs. Igualmente, es provisional. El ritmo de investigación actual en el campo invita a pensar que nuevas y mejoras ideas aparecerán en breve, junto con dificultades que habíamos pasado por alto y a las que habrá que atender.

Referencias

Abad, F. J., Olea, J., Aguado, D., Ponsoda, V., y Barrada, J. R. (2010). Deterioro de parámetros de los items en tests adaptativos informatizados: Estudio con eCAT. *Psicothema, 22*, 340-347.

Ariel, A., Veldkamp, B. P., y van der Linden, W. J. (2004). Constructing rotating item pools for constrained adaptive testing. *Journal of Educational Measurement, 41*, 345-359.

Ban, J., Hanson, B. A., Wang, T., Yi, Q., y Harris, D. J. (2001). A comparative study of on-line pretest item calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 191-212.

Ban, J., Hanson, B. A., Yi, Q., y Harris, D. J. (2002). Data sparseness and on-line pretest item calibration-scaling methods in CAT. *Journal of Educational Measurement, 39*, 207-218.

Baker, F. B. y Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Barrada, J. R., Abad, F. J., y Olea, J. (2011). Varying the valuating function and the presentable bank in computerized adaptive testing. *Spanish Journal of Psychology, 14*, 500-508.

Barrada, J. R., Abad, F. J., y Veldkamp, B. P. (2009). Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. *Psicothema, 21*, 318-325.

Barrada. J. R., Mazuela, P., y Olea, J. (2006). Maximum information stratification method for controlling item exposure in computerized adaptive testing. *Psicothema, 18*, 156-159.

Barrada, J. R., Olea, J., y Abad, F. J. (2008). Rotating item banks versus restriction of maximum exposure rates in computerized adaptive testing. *Spanish Journal of Psychology, 11*, 618-625.

Barrada, J. R., Olea, J., y Ponsoda, V. (2007). Methods for restricting maximum exposure rate in computerized adaptative testing. *Methodology, 3*, 14-23.

Barrada, J. R., Olea, J., Ponsoda, V., y Abad, F. J. (2008). Incorporating randomness to the Fisher information for improving item exposure control in CATs. *British Journal of Mathematical and Statistical Psychology, 61*, 493-513.

Barrada, J. R., Olea, J., Ponsoda, V., y Abad, F. J. (2009). Item selection rules in computerized adaptive testing: Accuracy and security. *Methodology, 5*, 7-17.

Barrada, J. R., Olea, J., Ponsoda, V., y Abad, F. J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement, 34*, 438-452.

Barrada, J. R., Veldkamp, B. P., y Olea, J. (2009). Multiple maximum exposure rates in computerized adaptive testing. *Applied Psychological Measurement, 33*, 58-73.

Birnbaum, A. (1968). Some latent ability models and their use in inferring an examinee's ability. En F. M. Lord y M. R. Novick (Eds.) *Statistical theories of mental test scores* (págs. 392-479). Reading, MA: Addison-Wesley.

Bock, R. D., y Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Buyske, S. (2005). Optimal design in educational testing. En M. P. F. Berger y W. K. Wong (Eds.), *Applied optimal designs* (págs. 1–16). New York: Wiley.

Chang, H. H.. (2004). Understanding computerized adaptive testing: From Robbins-Monro to Lord and beyond. In David Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (págs. 117-133). Thousand Oaks, CA: Sage Publications.

Chang, H. H., y Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

Chang, H. H., y Ying, Z. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211-222.

Chang, H. H., y Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika, 73*, 441-450.

Chang, H.-H., y Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics, 37*, 1466-1488.

Chang, H. H., y Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67*, 387-398.

Chang, S. W., y Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40*, 71-103.

Chang, S. W., y Harris, D. J. (2002, Abril). *Redeveloping the exposure control parameters of CAT items when a pool is modified*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans LA.

Chang, Y. C. I., y Lu, H. Y. (2010). Online calibration via variable length computerized adaptive testing. *Psychometrika, 75*, 140-157.

Chen, S. Y. (2010). A procedure for controlling general test overlap in computerized adaptive testing. *Applied Psychological Measurement, 34*, 393-409.

Chen, S. Y., y Ankenmann, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement, 41*, 149-174.

Chen, S. Y., Ankenmann, R. D., y Chang, H. H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24*, 241-255.

Chen, S. Y., Ankenmann, R. D., y Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40*, 129-145.

Chen, S. Y., y Doong, S. H. (2008). Predicting item exposure parameters in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 61*, 75-91.

Chen, S. Y., y Lei, P. W. (2005). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement, 29*, 204–217.

Chen, S., Lei, P., y Liao, W. (2008). Controlling item exposure and test overlap on the fly in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 61*, 471-492.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*, 619-632.

Cheng, Y., y Chang. H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62*, 369-383.

Cheng, Y., Chang, H. H., Douglas, J., y Guo, F. (2009). Constraint-weighted a-stratification for computerized adaptive testing with nonstatistical constraints. *Educational and Psychological Measurement, 69*, 35-49.

Choi, S. W., y Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33*, 419-440.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. New Jersey, Lawrence Erlbaum Associates.

Davey, T., y Parshall, C. G. (1995, Abril). New algorithms for item selection and exposure control with adaptive testing. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Davey, T., y Nering, N. (2002). Controlling item exposure and maintaining item security. En C. N. Mills, M. T. Potenza, J. J. Fremer, y W. C. Ward, (Eds). *Computer-based testing: Building the foundation for future assessments* (págs. 165-191). Mahwah, NJ: Lawrence Erlbaum.

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement, 16*, 327-343

Dodd, B. G. (1990) The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement, 14*, 355-366.

Embretson, S. E., y Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Finkelman, M., Nering, M. L., y Roussos, L. A. (2009). A conditional exposure control method for multidimensional adaptive testing. *Journal of Educational Measurement, 46*, 84-103.

Georgiadou, E., Triantafillou, E., y Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment, 5*(8). Accedido al 26 de Junio de 2007, desde http://www.jtla.org.

Hambleton, R. K., y Swaminathan, H. (1985). *Item response theory: Principles and applications*. Hingham, MA: Kluwer.

Hau, K. T., y Chang, H. H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38*, 249-266.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*, 44-52.

Hetter, R. D., y Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. En W. A. Sands, B. K. Waters, y J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (págs. 141-144). Washington DC: American Psychological Association.

Honan, W. H. (1995, 4 de Enero). Computer admissions test to be given less often. *The New York Times*, pág. A16. Accedido al 11 de Noviembre de 2009, desde http://www.nytimes.com/1995/01/04/us/computer-admissions-test-to-be-given-less-often.html.

Kingsbury, G. G., y Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4*, 241-261.

Lei, P., Chen, S., y Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement, 43*, 245-264.

Leung, C. K., Chang, H. H., y Hau, K. T. (2005). Computerized adaptive testing: a mixture item selection approach for constrained situations. *British Journal of Mathematical and Statistical Psychology, 58*, 239-257.

Li, Y. H., y Schafer, W. D. (2003, Abril). *The effect of item selection methods on the variability of CAT's ability estimates when item parameters are contaminated with measurement errors*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Li, Y. H., y Schafer, W. D. (2005). Increasing the homogeneity of CAT's item-exposure rates by minimizing or maximizing varied target functions while assembling shadow tests. *Journal of Educational Measurement, 42*, 245-269.

López-Cuadrado, J., Pérez, T. A., Vadillo, J. A., y Gutiérrez, J. (2010). Calibration of an item bank for the assessment of Basque language knowledge. *Computers & Education, 55*, 1044-1055.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*, 95-100.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*, 157-162.

Luecht, R.M. (1996). Multidimensional computer adaptive testing. *Applied Psychological Measurement, 20*, 389–404.

Luecht, R. M. (2005). Some useful cost-benefit criteria for evaluating computer-based test delivery models and systems. *Journal of Applied Testing Technology, 7*(2). Accedido el 24 de Julio de 2009, desde http://www.testpublishers.org/journal.htm.

Luecht, R. M., y Nungester R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*, 229-249.

Mills, C. N., y Steffen, M. (2000). The GRE computer adaptive test: Operation issues. En W. J. van der Linden y C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (págs. 75-100). Boston: Kluwer Academic Press.

Mills, C. N., y Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing, *Applied Measurement in Education, 9*, 287-304.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-195.

Mulder, J., y van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika, 74*, 273-296.

Muñiz, J. (1997). *Introduccion a la teoria de respuesta a los items*. Madrid: Piramide.

Nering M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*, 115-127.

Olea, J., Abad, F. J., Ponsoda, V., y Ximénez, M. C. (2004). Un test adaptativo informatizado para evaluar el conocimiento de inglés escrito: Diseño y comprobaciones psicométricas. *Psicothema, 16*, 519-525.

Olea, J. y Ponsoda, V. (2001). *Tests adaptativos informatizados*. Madrid: UNED.

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

Rebollo, P., García-Cueto, E., Zardaín, P. C., Cuervo, J., 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*, 241-251.

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

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.

Rulison, K. L., y Loken. E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement, 33*, 83-101.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-354.

Segall, D. O. (2004). A sharing item response theory model for computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 29*, 439-460.

Stern, E. B., y Havlick, L. (1986). Academic misconduct: Results of faculty and undergraduate student surveys. *Journal of Allied Health, 15*, 139-142.

Stocking, M. L., y Lewis, C. L. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57-75.

Stocking, M. L., y Lewis, C. L. (2000). Methods of controlling the exposure of items in CAT. En W. J. van der Linden y C. A. W. Glas (Eds.) *Computerized adaptive testing: Theory and practice* (pp. 163-182). Dordrecht, The Netherlands: Kluwer Academic.

Stocking, M. L, y Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277-292.

Sympson, J. B., y Hetter, R. D. (1985, Octubre). Controlling item-exposure rates in computerized adaptive testing. En *Proceedings of the 27th annual meeting of the Military Testing Association* (págs. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measuremen, 69*, 778-793.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201-216.

van der Linden, W. J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement, 23*, 21-29.

van der Linden, W.J. (2000). Constrained adaptive testing with shadow tests. En W.J. van der Linden y C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (págs. 27-52). Boston, MA: Kluwer Academic Publishers.

van der Linden. W. J. (2003). Some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 28*, 249-265.

van der Linden, W. J. (2005). A comparison of item-selection method for adaptive tests with content constraints. *Journal of Educational Measurement, 45*, 283-302.

van der Linden, W. J., y Glas, C. A. W. (Eds.) (2000). *Computerized Adaptive Testing. Theory and Practice*. Boston, MA: Kluwer Academic Publishers.

van der Linden, W. J., y Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259-270.

van der Linden, W. J., y Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*, 273-291.

van der Linden, W. J., y Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics 32,* 398-418.

Veerkamp, W. J. J., y Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*, 203-226.

Wainer, H. (2000a). CATs: Whither and whence. *Psicológica, 21*, 121-133.

Wainer, H. (2000b). Rescuing computerized testing by breaking Zipf's law. *Journal of Educational and Behavioral Statistics, 25*, 203-224.

Wainer, H., Bradlow, E. T., y Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*, 17-27.

Xu, X. L., y Douglas, J. (2006). Computerized adaptive testing under non-parametric IRT models. *Psychometrika, 71*, 121-137.

Yi, Q., Zhang, J., y Chang, H. H. (2008). Severity of organized item theft in computerized adaptive testing: A simulation study. *Applied Psychological Measurement, 32,* 543-558.

Zwick, R. (2000). The assessment of differential item functioning in computer adaptive tests. En W. J. van der Linden y C. A. W. Glas (Eds.), *Computerized Adaptive Testing. Theory and Practice* (págs. 221-243). Boston, MA: Kluwer Academic Publishers.

(Artículo recibido: 07-09-2010; revisión: 18-11-2010; aceptado: 20-11-2010)