Psicothema 2006. Vol. 18, nº 4, pp. 828-834
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

METODOLOGÍA

Estrategias de selección de ítems en un test adaptativo informatizado para la evaluación de inglés escrito

Juan Ramón Barrada, Julio Olea, Vicente Ponsoda y Francisco J. Abad
Universidad Autónoma de Madrid

e-CAT es un Test Adaptativo Informatizado para la evaluación del nivel de conocimiento del inglés escrito, que implementa la regla de selección de ítems más comúnmente empleada: el criterio de máxima información de Fisher. Algunos de los problemas asociados a este criterio de selección repercuten negativamente en la precisión de las estimaciones y en la seguridad del banco de ítems. En el presente trabajo se compara mediante simulación el rendimiento de esta regla con otras dos: la selección del ítem con máxima información de Fisher por intervalo de Veerkamp y Berger (1997) y una nueva regla, denominada como «máxima información de Fisher por intervalo con media geométrica». En general, este nuevo criterio de selección de ítems proporciona menor error de medida y menores tasas de solapamiento de ítems. Parece, pues, recomendable, al permitir obtener mejoras simultáneas en la calidad de las estimaciones y en el mantenimiento de la seguridad del banco de ítems en que se sustenta e-CAT.

Item selection rules in a Computerized Adaptive Test for the assessment of written English. e-CAT is a Computerized Adaptive Test for the evaluation of written English knowledge, using the item selection rule most commonly employed: the maximum Fisher information criterion. Some of the problems of this criterion have a negative impact in the estimation accuracy and in the item bank security. In this study, the performance of this item selection rule is compared, by means of simulation, with two other rules: selecting the item with maximum Fisher information in an interval (Veerkamp y Berger, 1997) and a new criterion, called «maximum Fisher information in an interval with geometric mean». In general, this new rule shows smaller measurement error and smaller item overlap rates. It seems, thus, recommendable, as it allows the simultaneous improvement of estimation accuracy and the maintenance of the item bank security of e-CAT.

La idea fundamental de un Test Adaptativo Informatizado (TAI) es emplear un ordenador y un modelo psicométrico de la Teoría de la Respuesta al Ítem (TRI) para presentar a las personas los ítems de un banco calibrado que resultan más apropiados para su nivel (Olea y Ponsoda, 2003; Ponsoda y Olea, 2003). Para ello se requiere fundamentalmente: a) un banco de ítems calibrado desde los desarrollos de un modelo de la TRI; b) un método estadístico para la estimación progresiva de los niveles de rasgo ($\hat{\theta}$); y c) un método para seleccionar progresivamente los ítems a presentar. A diferencia de los tests convencionales, en los TAIs no todos los evaluados reciben idénticos ítems: el ítem *(q+1)*-ésimo presentado varía según cuál sea el patrón de respuesta a los *q* ítems anteriores. Con ello se pretende conseguir estimaciones del nivel de rasgo más precisas que mediante la aplicación de tests lineales, dado que será necesario presentar un número menor de ítems (precisamente los que más contribuyen a reducir la incertidumbre sobre el nivel de rasgo que tiene la persona) para conseguir niveles semejantes de fiabilidad; en contextos de evaluación psicológica o educativa a gran escala, los TAIs pueden suponer mejoras importantes en las condiciones de aplicación y, a medio o largo plazo, un abaratamiento de costes.

La aplicación de este tipo de tests resulta ya casi tradicional en contextos de evaluación psicológica y educativa en países como Estados Unidos y Holanda, donde se aplican fundamentalmente para objetivos de evaluación del conocimiento, acreditación profesional y selección de personal. En España, al menos dos tests adaptativos se encuentran ya operativos: uno, denominado TRAS-I (Ru-

Fecha recepción: 14-11-05 • Fecha aceptación: 27-3-06
Correspondencia: Juan Ramón Barrada
Facultad de Psicología
Universidad Autónoma de Madrid
28049 Madrid (Spain)
E-mail: juanra.barrada@uam.es

bio y Santacreu, 2004), evalúa la capacidad de razonamiento secuencial e inductivo; el otro, denominado e-CAT (Olea, Abad, Ponsoda y Ximénez, 2004), evalúa el conocimiento del inglés escrito.

En el presente trabajo se realizan propuestas para mejorar la regla de selección de ítems implementada en este último TAI, para intentar de forma simultánea mejorar la precisión de las estimaciones y la seguridad del banco de ítems.

Selección de ítems mediante máxima información: descripción y limitaciones

Los dos TAIs citados, como la mayoría de los que se aplican en otros países, emplean el criterio de máxima información de Fisher (MIF) para seleccionar el mejor ítem para un evaluando. Supongamos que un evaluando ha respondido ya a *q* ítems del TAI; en este momento, mediante los procedimientos estadísticos (máximo-verosímiles o bayesianos) aplicados al modelo de TRI en uso, se estima un nivel de rasgo provisional ($\hat{\theta}_q$) con una precisión concreta (*Se*). Debemos entonces seleccionar el siguiente *q+1* ítem. El método MIF consiste en elegir el ítem que proporciona más información para el nivel provisional $\hat{\theta}_q$. Si, como es el caso en la prueba e-CAT, el banco de ítems está calibrado mediante el modelo logístico de 3 parámetros, la probabilidad de acertar un ítem se formula como:

$$P(\theta) = c + \frac{1-c}{1+e^{-1.7a(\theta-b)}} \quad (1)$$

donde:

a es el parámetro de discriminación;
b es el parámetro de dificultad;
c es el parámetro de pseudo-azar.

Para este modelo, la función de información de Fisher de un ítem puede calcularse mediante la ecuación 2 (Lord, 1977).

$$I(\theta) = \frac{2.89a^2(1-c)}{(c+e^{1.7a(\theta-b)})(1+e^{-1.7a(\theta-b)})^2} \quad (2)$$

En la ecuación puede verse que la información que aporta un ítem para un determinado nivel de rasgo será tanto mayor cuanto: a) mayor sea su parámetro de discriminación; b) menor sea su parámetro de pseudo-azar; y c) menor sea el valor absoluto de la diferencia entre el nivel de rasgo y el parámetro de dificultad. Según aumenta el valor de la función de información de Fisher para $\hat{\theta}$ se reduce el error típico de medida de la estimación (véase Muñiz, 1997, p. 125).

Ahora bien, esta RSI puede presentar limitaciones en dos de los objetivos a optimizar en los TAIs: la precisión y la seguridad del banco de ítems.

Los problemas asociados a la precisión tienen que ver con seleccionar los ítems más informativos para un nivel de rasgo provisional que puede estar alejado del nivel de rasgo verdadero de la persona. Por ejemplo, al comienzo del TAI pueden presentarse ítems altamente informativos para un nivel $\hat{\theta}$ que dista bastante del nivel de rasgo auténtico de la persona (θ), con lo que estaríamos seleccionando (y, por tanto, «gastando») ítems que no son los que más contribuyen a reducir la incertidumbre sobre este verdadero nivel. Además, la información de Fisher puede ser descrita como la capacidad de un ítem para discriminar entre dos puntos adyacentes de nivel de rasgo, es decir, para diferenciar entre $\hat{\theta}$ y un valor muy próximo. Sin embargo, para una eficiente estimación de θ, es deseable discriminar no únicamente entre niveles de rasgo próximos, sino también entre niveles distantes, especialmente en los primeros momentos de la administración del test (Chang y Ying, 1996).

La segunda limitación de la selección de ítems mediante MIF se refiere al control de exposición de los ítems. Con esta RSI unos pocos ítems, usualmente aquellos con mayor parámetro de discriminación, se presentan a una elevada proporción de examinados, mientras que una parte sustancial del banco apenas es utilizado. Esto supone un doble inconveniente. Por un lado, puede darse un problema de seguridad con los ítems sobreexpuestos: su contenido puede ser conocido previamente a su administración, dado que es alta la probabilidad de que dos evaluados reciban algunos ítems similares, con lo que dejarían de servir para la estimación del nivel de rasgo e introducirían error en la evaluación. Por otro lado, los ítems infraexpuestos representan un derroche económico, puesto que la inversión para su desarrollo no es recuperada.

En las primeras aplicaciones de e-CAT se ha comprobado que, si establecemos una tasa máxima de exposición de .25, casi la mitad de los ítems se acercan mucho a dicha tasa, lo que puede representar un problema importante, dado que esta prueba se aplica normalmente, a través de la web, a muestras numerosas de candidatos a determinados puestos de trabajo.

Para intentar minimizar el primero de los problemas descritos, Veerkamp y Berger (1997) propusieron, como alternativa a MIF, utilizar la función de Fisher por intervalo (MIF-I) como criterio de selección, tal y como se indica en la fórmula 3.

$$\max_{i \in B_n} \int_{\theta\,\min}^{\theta\,\max} I(\theta) d(\theta) \quad (3)$$

Según este criterio se seleccionaría aquel elemento, perteneciente al banco de ítems no presentados todavía a ese evaluado (B_n), que aporta la máxima información de Fisher para un intervalo de niveles de rasgo, y no para un valor concreto. Los límites del intervalo (θ_{min}, θ_{max}) se ajustan con cada nuevo ítem administrado para cubrir el área en la que es máximamente probable encontrar θ dado $\hat{\theta}$ e $I(\hat{\theta})$. A medida que avanza el test y se incrementa la información disponible para el nivel de rasgo estimado, el intervalo se va reduciendo, convergiendo de este modo MIF-I hacia MIF.

No parecen concluyentes los resultados que ofrecen sobre precisión las reglas MIF-I e MIF. Mientras que los datos de Veerkamp y Bergen (1997) parecen no indicar diferencias entre MIF y MIF-I, los de Chen, Ankenmann y Chang (2000) muestran una mayor precisión para MIF-I. Sin embargo, en el estudio de Barrada, Olea y Ponsoda (2004), parece que la regla con menor error de medida es MIF. Es posible que las características psicométricas de los bancos de ítems influyan en estos resultados aparentemente contradictorios, por lo que será necesaria estudiar la tendencia concreta que se consigue con los parámetros estimados en el banco que soporta el sistema e-CAT.

En estos trabajos no se ha estudiado los efectos sobre el control de la exposición, por lo que no sabemos si esta RSI podría ofrecer mejoras para la seguridad del banco objeto de estudio, en comparación con MIF. La lógica de selección de MIF-I permite suponer que pudiera ser así. Para un banco de ítems calibrado según el modelo de 3 parámetros, es posible encontrar ítems de baja tasa de ex-

posición según MIF que serían seleccionados más veces cuando se consideran regiones amplias de θ. En todo caso, es de prever un impacto limitado en el control de la exposición al aplicar MIF-I. El estudio de Chen et al. (2000) evaluó la coincidencia entre los ítems seleccionados según MIF y MIF-I. Estos autores encontraron tasas de coincidencia, condicionadas a θ, de entre .75 y .90 para 20 ítems presentados; significa esto que dos evaluados de igual θ, cada uno con distinta RSI, compartirán, como valor esperado, entre 15 y 18 ítems.

Una nueva RSI: información de Fisher por intervalo con media geométrica

La regla MIF-I tiene el inconveniente adicional de que puede seleccionar ítems distintos dependiendo de la variabilidad de la función de información. Imaginemos, por ejemplo, que un ítem *m* tiene una función de información con elevado grado de apuntamiento (alta curtosis) y parámetro *b* alejado del valor $\hat{\theta}_q$, mientras que la de otro ítem (*l*) tiende a ser platicúrtica (baja curtosis) y su parámetro a y $\hat{\theta}_q$ son próximos. Si el intervalo de integración que definimos es amplio puede ocurrir que el ítem *m* sea seleccionado en lugar del ítem *l*, si bien la información proporcionada por el ítem *m* para $\hat{\theta}_q$ sería menor que la ofrecida por *l*. Veerkamp y Berger (1997) han mostrado que, para intervalos amplios de integración, el funcionamiento de MIF-I depende básicamente del parámetro *a*, no de la distancia entre el parámetro *b* y el nivel de rasgo estimado.

Desarrollando una RSI que incorporara, directamente o indirectamente, la variabilidad en la función de información, ponderándola de forma inversa, sería posible la reducción o eliminación de casos como el descrito. Esto puede hacerse mediante el uso de la media geométrica que, para una serie de datos de *z* elementos, es la raíz *(1/z)*-ésima del multiplicatorio de los mismos. Para conjuntos de datos de igual media aritmética, el de mayor media geométrica será el que tenga menor varianza. Por ejemplo, los conjuntos (4, 6) y (1, 9) tienen, ambos, una media aritmética de 5, mientras que la media geométrica del primero es 4.9 y la del segundo es 3.

La nueva RSI que presentamos, que incorpora estas ideas, la hemos llamado «información de Fisher por intervalo con media geométrica» – MIF-IG. Esta RSI, muy próxima en su fundamento a MIF-I, busca superar alguna de sus limitaciones. La operativización de esta RSI requiere de una pequeña modificación en la ecuación 3. Ya no es posible integrar en un intervalo, sino que ha de emplearse el multiplicatorio, tal y como se muestra en la ecuación 4.

$$\max_{i\in B_n}\left(\prod_{j=0}^{k} V(\theta_j)W(\theta_j,x,g)\right)^{\frac{1}{k+1}} \quad \theta_j = \theta_{\min} + \frac{j(\theta_{\max}-\theta_{\min})}{k} \quad k = 80 \tag{4}$$

Creemos que con este nuevo criterio de selección pueden producirse dos efectos comparativos sustanciales: reducir la exposición de ítems de alto parámetro *a* y favorecer la selección de ítems de bajo parámetro *a*. Ambos efectos nos hacen pensar que MIF-IG ofrezca un mejor control de la exposición. Al evaluar la información en un intervalo, esperamos mejoras también en la precisión.

Estudio de simulación

Los objetivos planteados pueden estudiarse mediante simulación, tomando como datos de partida los parámetros de los ítems estimados empíricamente en e-CAT, que serán considerados como parámetros verdaderos.

Método

El banco está formado por un total de 197 ítems de opción múltiple, con 4 categorías de respuesta de las que sólo 1 es correcta. Los valores de media, desviación típica, máximo y mínimo para los parámetros *a*, *b* y *c* son (1.3, .32, 2.2, .43) (.23, 1, 3.42, -2.71) y (.21, .03, .29, .11), respectivamente. Información sobre el proceso de construcción, el diseño de anclaje, datos de ajuste al modelo y función de información puede consultarse en Olea, Abad y Ponsoda (2002), y Olea, Abad, Ponsoda y Ximénez (2004).

Como procedimiento de arranque del TAI se estableció un nivel de rasgo inicial ($\hat{\theta}_0$) elegido al azar dentro del intervalo (-0.5, 0.5). El primer intervalo fue fijado en ($\hat{\theta}_0$ - 3, $\hat{\theta}_0$ + 3).

Es conocido que la estimación máximo-verosímil no proporciona valores reales cuando el patrón de respuestas es constante, es decir, todo aciertos o todo errores. Por ello, hasta el momento en el que había un mínimo de variabilidad en las respuestas simuladas, θ era asignada mediante el método propuesto por Dodd (1990). En el caso de que el patrón de respuestas sean aciertos, $\hat{\theta}$ se incrementa en $(b_{max} - \hat{\theta})/2$; de ser todo errores, $\hat{\theta}$ se reduce en $(\hat{\theta} - b_{min})/2$. En el momento en que, para un sujeto simulado, aparezca variabilidad en las respuestas o se alcance el final del test, se aplica la estimación máximo-verosímil, forzando que $\hat{\theta}$ se encuentre dentro del intervalo [-4, 4].

Se generaron 50.000 niveles de rasgo verdaderos (sujetos simulados), extraídos al azar de una población N(0, 1), obteniendo sus respuestas en un total de 20 ítems. Resultados anteriores han mostrado que, a partir de esta longitud del TAI, la diferencia en precisión entre las reglas derivadas del criterio de máxima información es ya muy reducida.

Las distintas RSIs se compararon respecto a las siguientes variables dependientes:

(a) RMSE relativa a la precisión de medida (ecuación 5).
(b) La tasa de solapamiento ($\hat{T}$) para evaluar la seguridad del banco de ítems (ecuación 6), según la fórmula desarrollada por Chen, Ankenmann y Spray (2003). La tasa de solapamiento es la proporción de ítems que comparten, como media, dos examinados.
(c) Los valores medios de los parámetros *a* y *c* de los ítems administrados, con el objeto de analizar el tipo de ítem que tiende a ser elegido por cada RSI.
(d) La correlación de las tasas de exposición de los ítems para cada par de RSIs, como indicativo de convergencia entre ellas.

$$RMSE = \left(\sum_{i=1}^{r}\left(\hat{\theta}_i - \theta_i\right)^2 \Big/ r\right)^{1/2} \tag{5}$$

$$\hat{T} = \frac{n}{q}S_{er}^2 + \frac{q}{n} \tag{6}$$

donde:

r es el número de réplicas-evaluados.

$\hat{T}$ es el estimador poblacional de la tasa de la solapamiento (Chen et al., 2003).

n es el tamaño del banco de ítems.
q es el número de ítems administrados.
S^2_{er} es la varianza de las tasas de exposición de los ítems.

En cuanto a la aplicación de las diferentes RSI, el procedimiento MIF puede ser evaluado directamente mediante la función de información de Fisher para $\hat{\theta}$, tal y como se indica en la ecuación 2. Para las otras dos RSIs, los valores de $\theta_{\min}$ y de $\theta_{\max}$ fueron establecidos para α igual 0.05, tal y como se indica en las ecuaciones 7 y 8.

$$\theta_{\min} = \max\left(\hat{\theta} - 3, \hat{\theta} - \frac{\Phi^{-1}(.975)}{\sqrt{I(\hat{\theta})}}\right) \quad (7)$$

$$\theta_{\max} = \min\left(\hat{\theta} + 3, \hat{\theta} + \frac{\Phi^{-1}(.975)}{\sqrt{I(\hat{\theta})}}\right) \quad (8)$$

donde Φ es la distribución normal estándar acumulativa.

La función criterio de MIF-I, desarrollada, puede expresarse así (Veerkamp y Berger, 1997):

$$\int_{\theta\,\min}^{\theta\,\max} I(\theta)d(\theta) = \frac{1.7a}{(1-c)}\left[c\ln\left(\frac{P(\theta_{\min})}{P(\theta_{\max})}\right) + P(\theta_{\max}) - P(\theta_{\min})\right] \quad (9)$$

Para MIF-IG, la función criterio no puede ser resuelta analíticamente. Aplicando unas sencillas transformaciones a la ecuación 4, el valor de MIF-IG fue calculado mediante la ecuación 10. La ecuación 4 y la ecuación 10 ofrecen la misma ordenación de ítems en el criterio, si bien el sumatorio de logaritmos resulta computacionalmente más sencillo.

$$\sum_{j=0}^{k} 1\ln\left[I(\theta_j)\right] \qquad \theta_j = \theta_{\min} + \frac{j(\theta_{\max} - \theta_{\min})}{k} \qquad k = 80 \quad (10)$$

Resultados

En la figura 1 pueden verse los diferentes valores del RMSE obtenidos para distintas longitudes del TAI. Como cabe esperar, los valores medios de RMSE se reducen a medida que se incrementa el número de ítems administrados. La diferencia en RMSE entre las tres RSIs se va reduciendo según se incrementa el número de ítems presentados. Para cualquier longitud del test, con MIF se obtiene un RMSE menor que con MIF-I. La posición relativa de MIF-IG en RMSE varía según el número de ítems presentados. Al comienzo del test, esta nueva RSI es la menos precisa. A partir del sexto ítem presentado, con MIF-IG se obtiene un RMSE menor que con MIF-I. A partir del noveno ítem, MIF-IG es la RSI más precisa, si bien las diferencias con MIF son reducidas. En condiciones realistas de aplicación de e-CAT, donde resulta impensable aplicar menos de 15 ítems, la nueva regla de selección de ítems propuesta proporciona un error de medida ligeramente inferior a las otras dos.

En la figura 2 se muestran las tasas de solapamiento para las distintas RSIs. Las diferencias entre ellas son mayores al inicio del TAI, reduciéndose según avanza la presentación de ítems. Incrementar la longitud del test tiene un efecto en la tasa de solapamiento diferente para cada RSI. En el caso de MIF, la menor tasa de solapamiento se obtiene al comienzo del test, estabilizándose a partir del tercer ítem. Para MIF-I, aumentar el número de ítems administrados implica reducir la tasa de solapamiento. Para MIF-IG, del ítem 2 al ítem 9 se reduce la tasa de solapamiento, incrementándose ligeramente a partir de ese momento. Independientemente de la longitud del test, MIF-I es la RSI con mayor solapamiento. Para las otros dos RSIs, la eficacia en solapamiento depende del número de ítems administrados: hasta presentar el ítem quinto, MIF tiene un menor solapamiento; a partir de ese momento, lo que coincide con criterios realistas de parada del TAI, la RSI más segura es MIF-IG.

En la figura 3 se presentan las medias de los parámetros *a* y *c* de los ítems administrados. MIF y MIF-I, a diferencia de MIF-IG, tienden a seleccionar al comienzo del test ítems con parámetros *a* claramente superiores a la media de este parámetro en el banco (1.3). Para esas dos reglas, a medida que aumenta el número de ítems administrados, y los parámetros con mayor valor en *a* han sido ya seleccionados, se presentan ítems con parámetros *a* progresivamente menores. Con MIF-IG, los ítems de alto valor en el

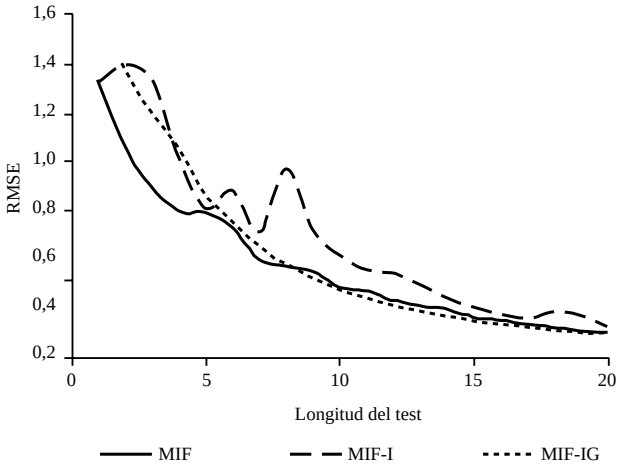

Figura 1. *RMSE según la longitud del test*

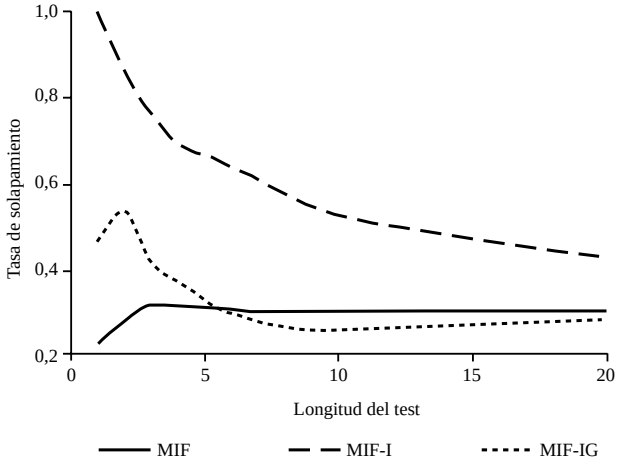

Figura 2. *Tasa de solapamiento según el número de ítems administrados*

parámetro *a* quedan disponibles para fases más avanzadas del test. Esto permite que el valor promedio al parámetro *a* cuando se administran más de 10 ítems sea mayor para MIF-IG que para las otras RSIs. Se observa que, superados los 10 ítems, desciende el valor medio del parámetro *a* administrado con MIF-IG. Para cualquier número de ítems administrados, MIF-I tiende a seleccionar ítems con mayores valores en el parámetro *a* que MIF.

Con respecto al parámetro *c*, las RSIs estudiadas tienden a seleccionar ítems con valores por debajo de .21, la media de este parámetro en el banco de ítems. La inestabilidad de las gráficas para los primeros ítems puede deberse a alguna característica específica de la distribución de parámetros en el banco empleado. Al igual que lo que ocurría con el parámetro *a*, pero en sentido inverso, según va avanzando el test y se van agotando los ítems con bajo valor en *c*, la media en este parámetro va incrementándose. La tendencia a seleccionar ítems de bajo valor en *c* es más acentuada para MIF-I y MIF-IG que para MIF.

Estas combinaciones de valores promedio de los parámetros *a* y *c* pueden explicar las diferencias en la tasa de solapamiento de las RSIs. MIF-I tiende a seleccionar, en mayor medida que MIF, ítems con parámetros de alto valor en *a* y bajo valor en el parámetro *c*, una combinación de valores infrecuente, lo que eleva la tasa de solapamiento de esta RSI. MIF-IG busca ítems de combinaciones de parámetros distintas según la fase del test en la que se encuentre. Al comienzo, ítems de bajo valor en el parámetro *a*, pasando a seleccionar cuando el test está más avanzado ítems de alto valor en *a*. Esta búsqueda de ítems distintos según la fase del test hace que la tasa de solapamiento sea menor que la encontrada para el resto de RSIs.

Se esperaba que las RSIs alternativas a MIF tendieran a seleccionar ítems distintos a MIF al comienzo del test, convergiendo a MIF según se incrementa el número de ítems administrados. Para evaluar esto se calculó, para cada par de RSIs, la correlación entre las tasas de exposición de los ítems, tal y como se muestra en la figura 4.

Tres son los resultados que pueden destacarse. Primero, las correlaciones, en líneas generales, se elevan según se incrementa el número de ítems administrados. Segundo, la coincidencia en el patrón de ítems seleccionados entre MIF y MIF-I ya es muy elevada incluso con pocos ítems administrados (igual que ocurre en el trabajo de Chen et al., 2000). Tercero, la RSI que más se diferencia de las demás en su patrón de selección de ítems es MIF-IG, como cabía esperarse a partir de los valores medios de los parámetros de los ítems seleccionados. La correlación entre las tasas de exposición para MIF y MIF-IG van desde -.07, cuando son 5 los ítems seleccionados, hasta .72, cuando ya son 20, marcadamente por debajo de la correlación de MIF-I con MIF. De hecho, hasta que no son 6 los ítems presentados, la correlación es negativa.

Discusión y conclusiones

El principal objetivo del presente estudio era intentar mejorar la precisión y seguridad del banco de ítems en que se sustenta e-CAT, un TAI para evaluar el nivel de inglés escrito que se aplica a través de Internet en contextos de selección de personal, mediante la modificación de la regla de selección de ítems. Para ello, se comparó mediante simulación el rendimiento de la regla de selección implementada actualmente en el sistema adaptativo (MIF) respecto a otra regla alternativa ya propuesta (MIF-I) y respecto a una propuesta original (MIF-IG). La idea fundamental del nuevo criterio de selección de ítems es establecer una medida por intervalo de la información, pero ponderando negativamente la variabilidad de la función de información.

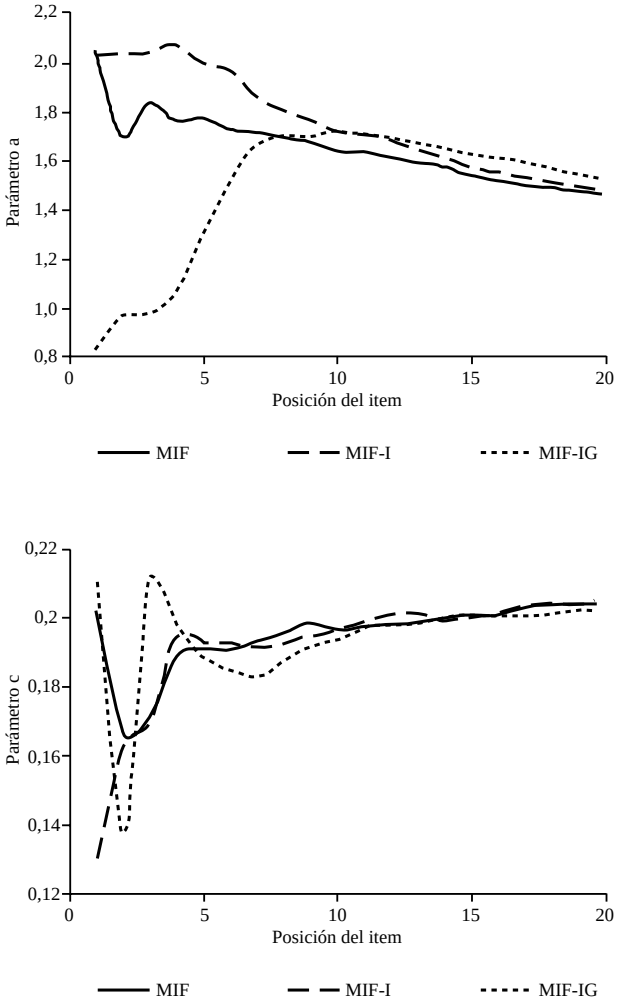

Figura 3. *Promedio de los parámetros a y c de los ítems administrados según la posición del ítem en el test*

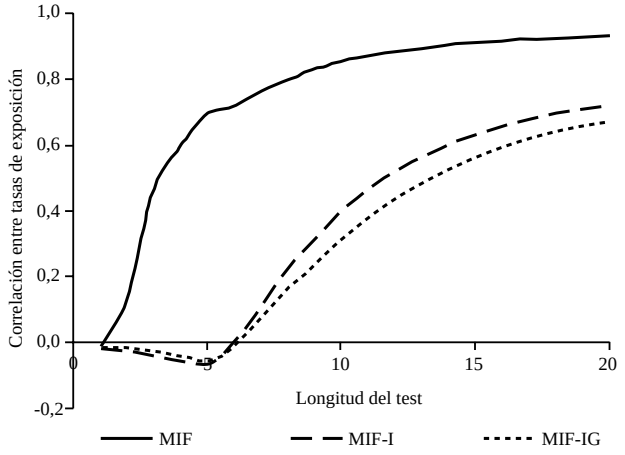

Figura 4. *Correlación entre las tasas de exposición para las distintas RSIs según la longitud del test*

El principal resultado obtenido es que, con el banco de ítems estudiado y en condiciones realistas de aplicación del TAI (entre 15 y 20 ítems como criterio de parada), la nueva regla de selección proporciona mejores niveles de precisión de las estimaciones y menor tasa de solapamiento entre ítems. Por encima de esa longitud del TAI, todas las RSIs tienden a converger hacia los mismos valores en las medidas de precisión y solapamiento.

Contrariamente a la presunción de varios autores (Chang y Ansley, 2003; Stocking y Lewis, 2000), parece que podemos obtener al mismo tiempo mejoras en precisión y seguridad para distintas RSIs. Al menos para el banco de ítems en que se sustenta e-CAT, es posible encontrar una RSI que, en comparación con las otras dos, mejora simultáneamente ambas variables. Aunque el presente estudio sirve para prever los efectos del cambio de RSI en e-CAT, debería estudiarse la constancia de estos resultados con bancos simulados de diferente longitud y con distintas distribuciones de parámetros.

Estudiando en detalle los efectos particulares de la regla MIF-IG, parece que sus ventajas (incluida la menor tasa de solapamiento obtenida) tienen que ver con las propiedades psicométricas de los ítems que selecciona: el comienzo del TAI se seleccionan ítems con bajo valor en el parámetro *a*, valor que se va incrementando hasta que se agotan los ítems de alta capacidad de discriminación.

Estos efectos pueden representar una ventaja adicional en el caso de que algunos evaluados conocieran de antemano algunos de los ítems del banco. Con pocos ítems administrados, el efecto de conocer de antemano el contenido y la respuesta de un ítem es mayor cuanto mayor es su parámetro *a* (Chang y Ying, 2002). Por eso, al igual que se hace en los métodos estratificados (Barrada, Mazuela y Olea, 2006; Chang y Ying, 1999), resulta conveniente que los primeros ítems administrados sean de bajo valor en el parámetro de discriminación.

Cabe resaltar además que, pese a tener formulaciones distintas, MIF e MIF-I muestran una elevada coincidencia en los ítems concretos que son seleccionados por ambos procedimientos, incluso en las fases iniciales del TAI (Chen et al., 2000). La única RSI con un patrón de selección diferenciado en los primeros ítems, que va convergiendo hacia MIF según avanza el test, es MIF-IG.

Consideramos que todavía puede mejorarse la nueva regla de selección propuesta. La aplicación de MIF-IG por la que hemos optado en este estudio, descrita en las ecuaciones 7, 8 y 10, podría modificarse para controlar la velocidad con la que esta regla converge a MIF. De este modo, podría buscarse que el crecimiento en el valor promedio del parámetro *a* de los ítems administrados fuera siempre creciente, no como ocurre con el método actual. Esta sería una posible futura línea de investigación.

Cabe destacar que todas las RSIs presentadas obtienen tasas de solapamiento superiores a los límites habitualmente considerados como aceptables (Way, 1998). Esto es debido a que el algoritmo no incorpora procedimientos eficientes para incrementar la infrautilización de ciertos ítems y reducir la sobreexposición de otros. Quedaría, por tanto, por investigar el efecto que tienen en las RSIs estudiadas métodos adicionales para un mejor control de la exposición, como los métodos de control de la tasa de exposición máxima (Sympson y Hetter, 1985; van der Linden y Veldkamp, 2004) o de control de la tasa de solapamiento (Chen y Lei, 2005).

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Ciencia y Tecnología (proyecto BSO2002-1485). El primero de los autores obtuvo el premio AEMCCO para Jóvenes Investigadores por la presentación de parte de este trabajo en el IX Congreso de Metodología de las Ciencias Sociales y de la Salud (Granada, septiembre del 2005).

Referencias

Barrada, J.R., Mazuela, P., y Olea, J. (2006). Maximum information stratification method for controlling item exposure in computerized adaptive testing. *Psicothema, 18*, 157-160.

Barrada, J.R., Olea, J., y Ponsoda, V. (2004). Reglas de selección de ítems en tests adaptativos informatizados. *Metodología de las Ciencias del Comportamiento, volumen especial*, 55-61.

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. (2002, abril). *To weight or not to weight - balancing influence of initial and later items in CAT*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

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.

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 Lei, P.W. (2005). Controlling ítem exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement, 29*, 204-217.

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

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

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

Olea, J., Abad, F.J., y Ponsoda, V. (2002). Elaboración de un banco de ítems, predicción de la dificultad y diseño de anclaje. *Metodología de las Ciencias del Comportamiento, volumen especial*, 427-430.

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. (2003). *Tests adaptativos informatizados*. Madrid: UNED.

Ponsoda, V,. y Olea, J. (2003). Adaptive and tailored testing (including IRT and non-IRT application). En R. Fernández-Ballesteros (ed.): *Encyclopedia of Psychological Assessment* (pp. 9-13). London: SAGE.

Rubio, V., y Santacreu, J. (2004). *TRAS-I: Test adaptativo informatizado para la evaluación del razonamiento secuencial y la inducción*. Madrid: TEA.

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.

Sympson, J.B., y Hetter, R.D. (1985, octubre). *Controlling ítem exposure rates in computerized adaptive testing.* Proceedings of the 27th annual meeting of the Military Testing Association (pp. 973-977). San Diego, CA.

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.

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

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