Psicothema 2010. Vol. 22, nº 2, pp. 340-347
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Deterioro de parámetros de los ítems en tests adaptativos informatizados: estudio con eCAT

Francisco José Abad, Julio Olea, David Aguado,Vicente Ponsoda y Juan Ramón Barrada*
Universidad Autónoma de Madrid y * Universidad Autónoma de Barcelona

En el presente trabajo se muestra el análisis realizado sobre un Test Adaptativo Informatizado (TAI) diseñado para la evaluación del nivel de inglés, denominado eCAT, con el objetivo de estudiar el deterioro de parámetros (parameter drift) producido desde la calibración inicial del banco de ítems. Se ha comparado la calibración original desarrollada para la puesta en servicio del TAI (N= 3224) y la calibración actual obtenida con las aplicaciones reales del TAI (N= 7254). Se ha analizado el Funcionamiento Diferencial de los Ítems (FDI) en función de los parámetros utilizados y se ha simulado el impacto que sobre el nivel de rasgo estimado tiene la variación en los parámetros. Los resultados muestran que se produce especialmente un deterioro de los parámetros *a* y *c*, que hay un importante número de ítems del banco para los que existe FDI y que la variación de los parámetros produce un impacto moderado en la estimación de θ de los evaluados con nivel de inglés alto. Se concluye que los parámetros de los ítems se han deteriorado y deben ser actualizados.

Item parameter drift in computerized adaptive testing: Study with eCAT. This study describes the parameter drift analysis conducted on eCAT (a Computerized Adaptive Test to assess the written English level of Spanish speakers). The original calibration of the item bank (N = 3224) was compared to a new calibration obtained from the data provided by most eCAT operative administrations (N = 7254). A Differential Item Functioning (DIF) study was conducted between the original and the new calibrations. The impact that the new parameters have on the trait level estimates was obtained by simulation. Results show that parameter drift is found especially for *a* and *c* parameters, an important number of bank items show DIF, and the parameter change has a moderate impact on high-level-English θ estimates. It is then recommended to replace the original estimates by the new set.

Un Test Adaptativo Informatizado (TAI) es una prueba de evaluación psicológica que se administra a través del ordenador y cuya característica distintiva es que la presentación de los ítems se realiza en función del nivel en la variable medida que va demostrando el evaluado (Olea y Ponsoda, 1996). Su uso se encuentra cada vez más extendido en contextos de evaluación psicológica y educativa, y pruebas ampliamente utilizadas en países como Estados Unidos u Holanda disponen habitualmente de versiones adaptativas. Es común que exámenes de licenciatura, certificación, acreditación o admisión se realicen de forma usual mediante TAIs. También en España disponemos ya de algunas pruebas adaptativas comercializadas, como el Test Informatizado para la Evaluación del Razonamiento Analítico, Secuencial e Inductivo, TRASI (Rubio y Santacreu, 2003) y el Test Adaptativo Informatizado para la Evaluación de Inglés, eCAT (Olea, Abad, Ponsoda y Ximénez, 2004).

Fecha recepción: 3/3/2009 • Fecha aceptación: 30/6/2009
Correspondencia: Francisco Jose Abad
Facultad de Psicología
Universidad Autónoma de Madrid
28049 Madrid (Spain)
e-mail: fjose.abad@uam.es

En este sentido, el estudio de procedimientos eficientes para el mantenimiento y la renovación de los bancos de ítems es uno de los desafíos principales a los que nos enfrentamos para que este tipo de pruebas de evaluación sigan aplicándose con éxito (Ponsoda, Hontangas, Olea, Revuelta, Abad y Ximénez, 2004). Es un problema bastante estudiado en Teoría de la Respuesta a los Ítems (TRI) que, a pesar de la teórica propiedad de invarianza de los estimadores, los parámetros de los ítems varíen en sucesivas aplicaciones del mismo test o del mismo TAI (v.gr., Donoghue y Isham, 1998; Glas, 2000; Wang y Kolen, 2001). Por ejemplo, para la construcción de eCAT, los datos para la calibración de los ítems se obtuvieron en unas condiciones que difieren apreciablemente de las condiciones en las que eCAT se aplica, en cuanto a (Olea et al., 2004): el tipo de formato (lápiz y papel vs informatizado), la posibilidad de omitir preguntas, de revisar las respuestas y el orden de presentación de los ítems según su dificultad (presentación fija frente a adaptativa). Y, por tanto, cabe pensar que algunos parámetros de los ítems se modifiquen en la aplicación adaptativa, haciendo necesaria una actualización de éstos.

El estudio de la modificación de parámetros requiere contrastar si la función de respuesta de un ítem es similar cuando se obtiene a partir de las respuestas iniciales de la muestra de calibración del banco (etapa pretest) y cuando se obtiene a partir de las respuestas

dadas en las sucesivas aplicaciones del TAI (etapa on-line). En tests fijos se han propuesto diversos procedimientos para estudiar si los parámetros de un ítem difieren entre dos grupos, lo que se conoce como Funcionamiento Diferencial del Ítem (FDI): basados en χ^2 para comprobar de forma simultánea si los parámetros de un ítem son estadísticamente equivalentes en dos aplicaciones distintas (Lord, 1980); obteniendo, mediante integración, el área delimitada por las funciones de respuesta obtenidas para un ítem en dos calibraciones (Raju, 1988); aplicando procedimientos de detección derivados del método de Mantel-Haenzel (Holland y Thayer, 1988), etc. Sin embargo, en el caso de los TAIs, la presentación adaptativa tiene algunos problemas particulares. En primer lugar, no puede emplearse la suma de aciertos como estimación de los niveles de rendimiento, tal como se hace muchas veces para obtener los niveles de FDI. En segundo lugar, los datos obtenidos en los TAIs son inherentemente incompletos, dado que diferentes evaluados responden a ítems distintos. Por tanto, no será raro encontrar cantidades importantes de ítems (los menos discriminativos) aplicados en pocas ocasiones y, por tanto, con poca cantidad de datos para realizar las nuevas estimaciones de sus parámetros. Y, en tercer lugar, dada su naturaleza adaptativa, existe una importante restricción del rango de θ disponible para cada ítem (los ítems fáciles los responden personas de bajo nivel, los difíciles las de alto nivel) cuando lo más adecuado sería disponer de una muestra amplia con distribución de rasgo uniforme (Stocking, 1990). Estas razones han llevado a proponer algunos métodos específicos para estudiar la variación de parámetros en los TAIs, unos derivados de procedimientos de estudio del FDI (Guo y Wang, 2003; Zwick, 2000; Zwick y Thayer, 2002) y otros específicos para detectar las consecuencias del conocimiento previo de los ítems (Glas, 2000).

Otra cuestión importante se refiere a la repercusión del cambio de los parámetros en la estimación de θ. En el contexto de los tests fijos de al menos 40 ítems, Wells, Subkoviak y Serlin (2002) encuentran que cuando *a* y *b* se incrementaban para el 20% de los ítems en 0,5 y 0,4 puntos, respectivamente, los efectos en la estimación de θ eran pequeños (menores de 0,14 logits). Sin embargo, es claro que el deterioro de los parámetros puede tener mayor efecto en un TAI que, usualmente, tendrá menor longitud. Por otro lado, en TAIs es más difícil evaluar el sesgo a nivel de test que se produce cuando los parámetros varían a través de los grupos. En tests fijos, se suele trabajar con el concepto de Funcionamiento Diferencial del Test (la diferencia en puntuación esperada para grupos igualados en el nivel de rasgo). En TAIs, cada persona responde a distintos tests, por lo que la puntuación esperada tiene poco valor práctico.

El objetivo del presente artículo es estudiar el deterioro de los parámetros de un TAI denominado eCAT (Olea et al., 2004) diseñado para la evaluación del nivel de dominio del inglés escrito de personas castellanoparlantes. Para ello: (a) se compara la calibración original (etapa pretest), realizada para la construcción de la prueba, con la calibración actual (etapa *on-line*) obtenida a través de las respuestas a eCAT en su utilización habitual; (b) se estudia la presencia de FDI al comparar ambas calibraciones; y (c) se analiza la repercusión que el cambio de los parámetros tiene en la estimación de θ, mediante una estrategia de simulación.

Se espera que debido a las diferentes condiciones de aplicación en ambas etapas (pretest y *on-line*) exista una modificación de los mismos en la aplicación adaptativa e informatizada y, por tanto, la actualización de los parámetros redundará en una mejora de las propiedades métricas de eCAT.

Método

Participantes

Muestra R. 3224 participantes, estudiantes de la Universidad Católica de Chile. El banco se dividió en cinco subtest para facilitar su aplicación. Cada subtest estuvo formado por 61 ítems (20 de anclaje y 41 propios de cada subtest). Cada participante respondió en un formato de lápiz y papel a un subtest asignado aleatoriamente y dispuso de 60 minutos para su cumplimentación. Es la muestra utilizada para la calibración inicial del banco de ítems de eCAT (calibración pretest).

Muestra F. 7254 personas, aspirantes a puestos de trabajo en diferentes organizaciones, que utilizan eCAT como parte de su proceso selectivo, y estudiantes de último curso de la Universidad Autónoma de Madrid, que lo realizan como parte de un programa para el desarrollo y evaluación de competencias transversales. Es la muestra utilizada para la calibración on-line.

Instrumentos

eCAT. Es un TAI diseñado para la evaluación del nivel de dominio del inglés escrito de personas castellano hablantes que se administra a través de internet a partir de conexión segura a una página WEB. Se compone de un banco de 197 elementos de elección múltiple con cuatro opciones de respuesta y tiene un formato de "cloze" (rellenar hueco). Información detallada acerca del desarrollo de los elementos se encuentra disponible en Olea, Abad y Ponsoda (2002). El algoritmo adaptativo tiene las siguientes características fundamentales (para una descripción detallada, véase Olea et al. 2004): a.) Procedimiento de arranque aleatorio: para comenzar la prueba, se elige un nivel de rasgo de una distribución normal truncada entre -1 y +1; b.) Estimación de θ: mediante máxima-verosimilitud o por una variación del método de Dodd mientras el patrón de respuestas sea constante (Dodd, 1990); c.) Selección de ítems: por máxima información, con algunas restricciones en el control de la exposición (p.e., la tasa máxima de exposición de un ítem es del 25%); d.) Procedimiento de parada: longitud fijada a 30 ítems.

El análisis de las aplicaciones adaptativas de eCAT muestra que el error típico medio es de 0,22, equivalente a una fiabilidad de 0,95. Para el 93% de los evaluados se consigue un error típico medio inferior a 0,30. El nivel medio de θ estimada es 0,67, lo que nos indica que el TAI se ha aplicado fundamentalmente a niveles medios-altos de comprensión del inglés escrito. Como promedio, se han empleado 20 segundos por ítem, lo que lleva a un tiempo medio de aplicación del test de 10 minutos. Respecto a la exposición de los ítems, 17 de ellos no han sido nunca seleccionados, ya que son ítems con bajo parámetro de dificultad, no ajustados a los niveles de inglés de los evaluados que responden a eCAT. La correlación entre los niveles θ estimados y el tiempo total empleado en el test fue de -0,166, una correlación inversa y significativa ($p<0,01$). El análisis de las correlaciones entre los parámetros de los ítems y sus tasas de exposición muestra, como era de esperar, que los ítems más expuestos son los de mayor parámetro *a* ($r_{TE,a} = 0,636$; $p<0,01$) y menor parámetro *c* ($r_{TE,c} = -0,345$; $p<0,01$) dado que son los ítems que aportan mayor nivel de información. Existe también una correlación significativa entre los parámetros de dificultad y las tasas de exposición ($r_{TE,b} = 0,356$; $p<0,01$), que tiene que ver directamente con los niveles de rasgo de la muestra que ha respondido a eCAT.

Procedimiento

Calibración pre-test. Para esta calibración se utilizaron los datos de la muestra *R*. Para la calibración del banco se realizó un diseño de anclaje en consonancia con la aplicación de subtests a la muestra *R*. Para los 197 elementos del banco se estimaron sus parámetros mediante el procedimiento de máxima verosimilitud marginal bayesiano implementado en BILOG (Mislevy y Bock, 1990). Las omisiones se trataron como respuestas fraccionalmente correctas. Para la distribución del nivel de habilidad se asumió una distribución normal (media= 0; desviación típica= 1). Se utilizó la opción FLOAT que permite estimar las medias para las distribuciones previas de los parámetros de los ítems. La distribución a priori inicial para los parámetros *a* era lognormal (media= 0,75; desviación típica= 0,12), para los parámetros *b*, normal (media= 0; desviación típica= 2) y para el parámetro *c* se utilizó una distribución beta (alpha= 76; beta= 226; es decir, con media 0,25, el recíproco del número de alternativas, y desviación típica 0,025). Posteriormente se estimaron las medias de la distribución previa para cada parámetro por un proceso iterativo. Las distribuciones previas finales (empíricamente estimadas) fueron log-normal para el parámetro *a* (media = 1,280, desviación típica = 0,205), normal para el parámetro b (media = 0,233, desviación típica = 2) y beta para el parámetro c (media = 0,207, desviación típica = 0,023). Información más detallada sobre el procedimiento de calibración, la distribución empírica de los parámetros de los ítems, la comprobación de la unidimensionalidad y la función de información del banco de ítems puede encontrarse en Olea et al. (2004) y en Abad, Olea, Ponsoda, Ximenez y Mazuela (2005).

Calibración on-line. Para esta calibración se utilizaron los datos de la muestra *F*. Debido a que en algunos estudios se ha mostrado que BILOG no alcanza la convergencia en el proceso de estimación de parámetros de los ítems cuando los datos provienen de la aplicación de TAIs (Harmes, Parshall y Kromrey, 2003) y cuando se establecen valores iniciales inadecuados para los parámetros (Pommerich y Segall, 2003), los ítems fueron calibrados a través del programa MULTILOG 7.0 (Thissen, Chen y Bock, 2003). Se utilizaron procedimientos de calibración concurrente, en el que las respuestas a ítems no aplicados a los sujetos se consideran como datos perdidos (Hanson y Béguin, 2002) y distribuciones previas similares a las utilizadas con BILOG en la calibración pretest. Para comprobar que las diferencias entre ambas calibraciones (pretest y *on-line*) no se debían al programa utilizado (BILOG y MULTILOG) se comprobó que los resultados eran muy similares cuando se aplicaban ambos programas a la misma muestra de respuestas pretest ($r_{aBILOG,aMULTILOG} = 0,983$; $r_{bBILOG,bMULTILOG} = 0,998$; $r_{cBILOG,cMULTILOG} = 0,950$). Para realizar esta comprobación, puesto que MULTILOG no permitía tratar las omisiones como fraccionalmente correctas, se sustituyeron estas por una respuesta aleatoria, estableciendo como probabilidad de acierto 0,25)

Puesto que los parámetros se han estimado en distintas muestras *R* y *F* que difieren en la distribución del rasgo, se requiere aplicar un proceso de equiparación para que los parámetros se encuentren en la misma métrica. En nuestro caso, se aplicó el procedimiento de Haebara (Haebara, 1980) donde para cada ítem se encuentran las constantes M_1 y M_2 que minimizan el criterio F:

$$F = \sum_{q=1}^{Q} g(\theta_q) \left(\sum_{i=1}^{n} \left(P(\theta_q; a_{iold}, b_{iold}, c_{iold}) - P(\theta_q; \frac{a_{inew}}{M_1}, M_1 b_{inew} + M_2, c_{inew}) \right)^2 \right) \quad (1)$$

Donde $g(\theta_q)$indica la probabilidad de tener $\theta = \theta_q$ asumiendo una distribución normal [N(0,1)] discretizada en *Q* puntos de cuadratura (en nuestro caso, $Q = 41$) y $P(\theta_q; a_{iold}, b_{iold}, c_{iold})$ y $P(\theta_q; \frac{a_{inew}}{M_1}, M_1 b_{inew} + M_2, c_{inew})$ son las probabilidades de acertar el ítem *i* estimadas según los parámetros originales (*old*) y según los nuevos parámetros (*new*) después de aplicar la transformación lineal a éstos.

Evaluación del FDI: Para analizar en qué ítems se produce FDI, se aplicó un procedimiento paramétrico basado en la TRI, usando el marco DFIT (Differential Functioning of Items and Test). Una vez equiparados los parámetros de TRI de ambos grupos (*R* y *F*) se obtuvo para cada ítem el indicador NCDIF.

$$NCDIF_i = \int_{-\infty}^{\infty} \left[P(x_i = 1 \mid \theta, g = F) - P(x_i = 1 \mid \theta, g = R) \right]^2 f(\theta \mid g = F) d\theta \quad (2)$$

Este indicador establece el promedio de las diferencias cuadráticas entre las CCIs de ambos grupos a través del rasgo (θ). El NCDIF es una medida muy similar a otras basadas en la TRI como el χ^2 de Lord o las medidas de área sin signo. El punto de corte para decidir que un ítem tiene FDI a partir de este indicador suele situarse en 0,006 (Raju, 1999). Estudios previos han mostrado que estos indicadores pueden funcionar mejor que otros procedimientos de TRI como el IRT-LR (Bolt, 2002). Para saber en qué dirección se producía la ventaja se calculó también la ventaja promedio:

$$V.P. = \int_{-\infty}^{\infty} \left[P(x_i = 1 \mid \theta, g = F) - P(x_i = 1 \mid \theta, g = R) \right] f(\theta \mid g = F) d\theta \quad (3)$$

Repercusión en θ estimada del cambio de parámetros: Siguiendo la propuesta de Guo y Wang (2003) se estudió, mediante simulación, la diferencia esperada en la θ estimada como función de los parámetros utilizados.

Se simularon las respuestas de 7254 sujetos de una distribución normal N(0,5, 1). La simulación se realiza utilizando los parámetros nuevos estimados en la muestra *F* (on-line) para la generación de las respuestas de los sujetos simulados y los parámetros originales estimados en la muestra *R* (pre-test) en la selección de ítems y la estimación de θ en cada paso. El criterio de parada se establece, al igual que en el TAI operativo, en 30 ítems. Posteriormente, se estudia el valor esperado de θ final estimada con los 30 ítems según los parámetros utilizados (*F* o *R*).

Resultados

Análisis de las calibraciones pre-test y on-line

En la tabla 1 se incluyen los datos descriptivos de los 3 parámetros estimados en ambas muestras. Los análisis se muestran para los 159 ítems que fueron aplicados en eCAT a más de 200 personas. Como puede apreciarse, en promedio se observan escasas discrepancias entre la media de los parámetros estimados en las diferentes condiciones.

Las correlaciones entre los parámetros estimados en las diferentes condiciones muestrales se detallan en la tabla 2. En relación al parámetro *b*, puede comprobarse una elevada relación lineal entre las estimaciones de cada parámetro realizadas en ambas condiciones muestrales. Las correlaciones para los parámetros *a* y *c* son algo menores, lo que indica que tiende a haber cambios en dichos parámetros. Las correlaciones cuando se se-

Tabla 1
Estadísticos descriptivos (Media, Desviación típica, Mínimo, Máximo) de los parámetros después de la equiparación, para los 159 ítems aplicados al menos 200 veces en eCAT.

	Media	Desviación Típica	Mínimo	Máximo
aR	1,39	0,29	0,91	2,20
aF	1,39	0,41	0,31	2,36
bR	0,35	1,02	-2,45	3,42
bF	0,32	0,95	-1,65	3,57
cR	0,20	0,03	0,11	0,29
cF	0,20	0,01	0,12	0,25

Tabla 2
Correlaciones entre los parámetros en las distintas muestras, para los 159 ítems aplicados al menos 200 veces en eCAT.

Correlación de Pearson entre los parámetros en las muestras R y F	
a	0,645
b	0,940
c	0,510

leccionan los 135 ítems aplicados a al menos 500 personas fueron muy similares (0,649, 0,936 y 0,506, respectivamente para *a, b y c*).

Tabla 3
Correlaciones entre el cambio en los parámetros *a, b* y *c* (cambio relativo y absoluto, muestras R y F) y la Tasa de Exposición (TE), para los 159 ítems aplicados al menos 200 veces en eCAT

		Tasa de Exposición (TE)
Cambio relativo	***aF-aR***	-0,184*
	bF-bR	-0,015
	cF-cR	0,130
Cambio absoluto	***|aF-aR|***	0,182*
	|bF-bR|	-0,090
	|cF-cR|	0,240**

** La correlación es significativa al nivel 0,01 (bilateral);

* La correlación es significativa al nivel 0,05 (bilateral)

Para comprobar si las variaciones en las estimaciones de los parámetros estaban relacionadas con las tasas de exposición de los ítems, se obtuvieron las diferencias en *a*, *b* y *c* entre las condiciones *R* y *F*; también se obtuvo el valor absoluto de dichas diferencias. En la tabla 3 se presentan las correlaciones entre estas variables y las tasas de exposición al comparar las muestras *R* y *F*.

En relación al parámetro *b* no existen relaciones estadísticamente significativas entre el tamaño del cambio, absoluto o relativo, y la tasa de exposición (p>0,05). Por tanto, no parece que haya habido difusión de los contenidos. A mayor exposición los ítems no se hacen más fáciles. Puede observarse que existe una

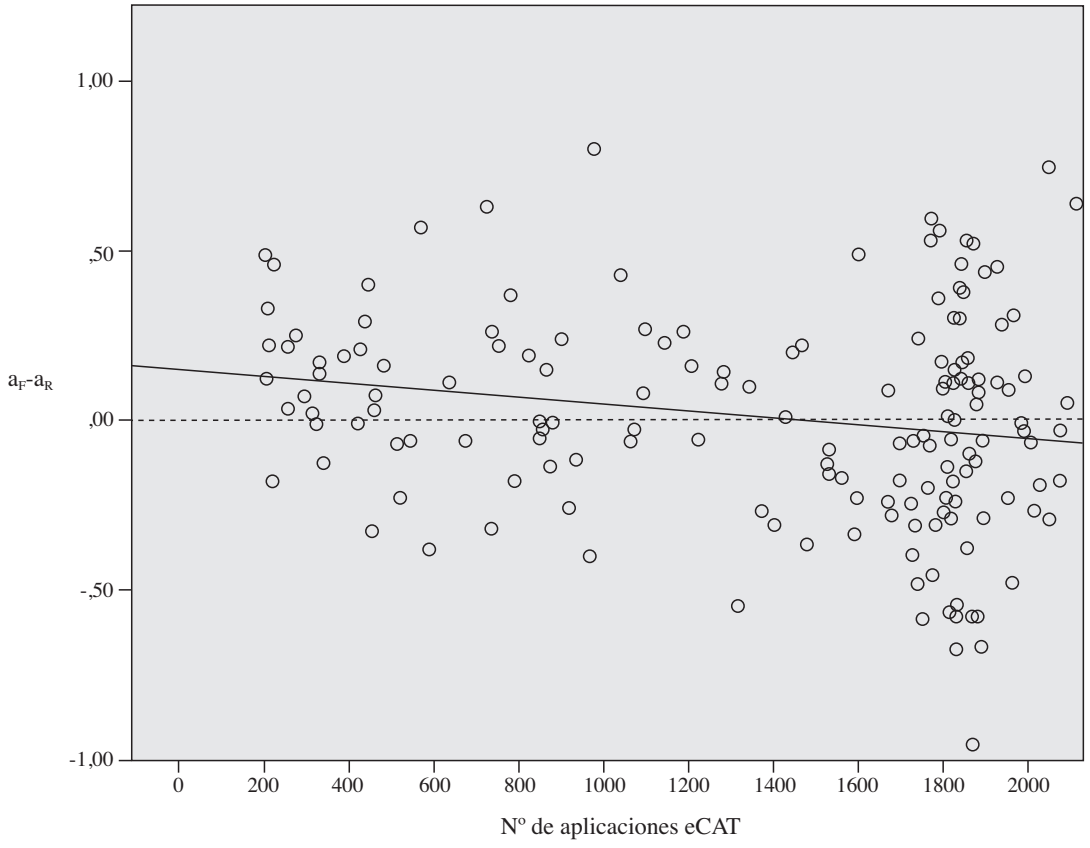

Figura 1. *Cambio relativo en a (a_F-a_R) como función del número de aplicaciones de eCAT*

pequeña relación entre las diferencias absolutas en los parámetros estimados y las tasas de exposición, tanto para el parámetro *a* ($r_{|aF-aR|,TE}$ = 0,182, p<0,05) como para el parámetro *c* ($r_{|cF-cR|,TE}$ = 0,240, p<0,01). Se puede observar cómo a mayor tasa de exposición, menores parámetros *a* ($r_{aF-aR,TE}$= -0,184; p<0,05). Podría pensarse que la correlación negativa encontrada para el parámetro *a* indica un deterioro de la calidad de un ítem dada su mayor tasa de exposición. Sin embargo, en un análisis más detallado, se observa que a medida que aumenta el número de aplicaciones la diferencia media entre los parámetros *a* en ambos formatos (*R* y *F*) se aproxima a 0 (véase figura 1).

El sesgo positivo en la estimación del parámetro *a* para los ítems menos aplicados puede explicarse por las características de la estimación bayesiana. Cuando no hay suficientes sujetos para estimar los parámetros, los valores de estos convergen a los especificados en las distribuciones previas. Los ítems menos expuestos son justamente los de menor parámetro *a* y por ello se observa un sesgo positivo (hacia la media de la distribución previa) cuando hay pocas aplicaciones. A medida que un ítem es más aplicado los cambios son mayores en términos absolutos, pero el sesgo medio se aproxima a 0.

En el parámetro *c*, el patrón es algo distinto. A medida que aumenta el número de aplicaciones, los valores de *c* en la muestra *F* tienden a tener más cambio en términos absolutos, pero no se observan tendencias sistemáticas de cambio (figura 2).

Evaluación del FDI

Se obtuvieron los valores del NCDIF para los ítems aplicados a más de 200 evaluados (159 ítems). Los resultados muestran que en un número considerable de ítems (59 ítems, 8 de los cuales fueron ítems de anclaje) ha habido cambios relevantes en los parámetros. La correlación entre tasas de exposición y valores de NCDIF y ventaja promedio no fueron significativas (r= 0,107 para NCDIF y r= 0,147 para la ventaja promedio; p>0,05), lo cual indica que los ítems con más variación en su CCI no son los que más veces se presentan en eCAT. La media de la ventaja promedio es próxima a cero (-0,003). Para el resto de los ítems contrastados (100 ítems) se puede considerar que sus parámetros son estables (NCDIF < 0,006).

También se realizó un análisis con el programa IRTLRDIF. Estos análisis sólo pudieron realizarse con 132 ítems, dado que el programa no convergía con algunos ítems. En todos los ítems contrastados se obtuvieron valores G^2 significativos, con lo que se concluye que en esos ítems habría habido un cambio significativo en algunos de los parámetros. Sin embargo, dados los tamaños muestrales, la significación estadística del FDI es poco ilustrativa.

Repercusión en θ estimada del cambio de parámetros

Los resultados muestran que el sesgo al estimar θ producido por utilizar los parámetros originales (en vez de los nuevos parámetros utilizados para generar las respuestas en la simulación) es pequeño aunque puede ser relevante en la práctica. La correlación entre ambos conjuntos de θ con la θ real es muy alta (r= 0,973 y

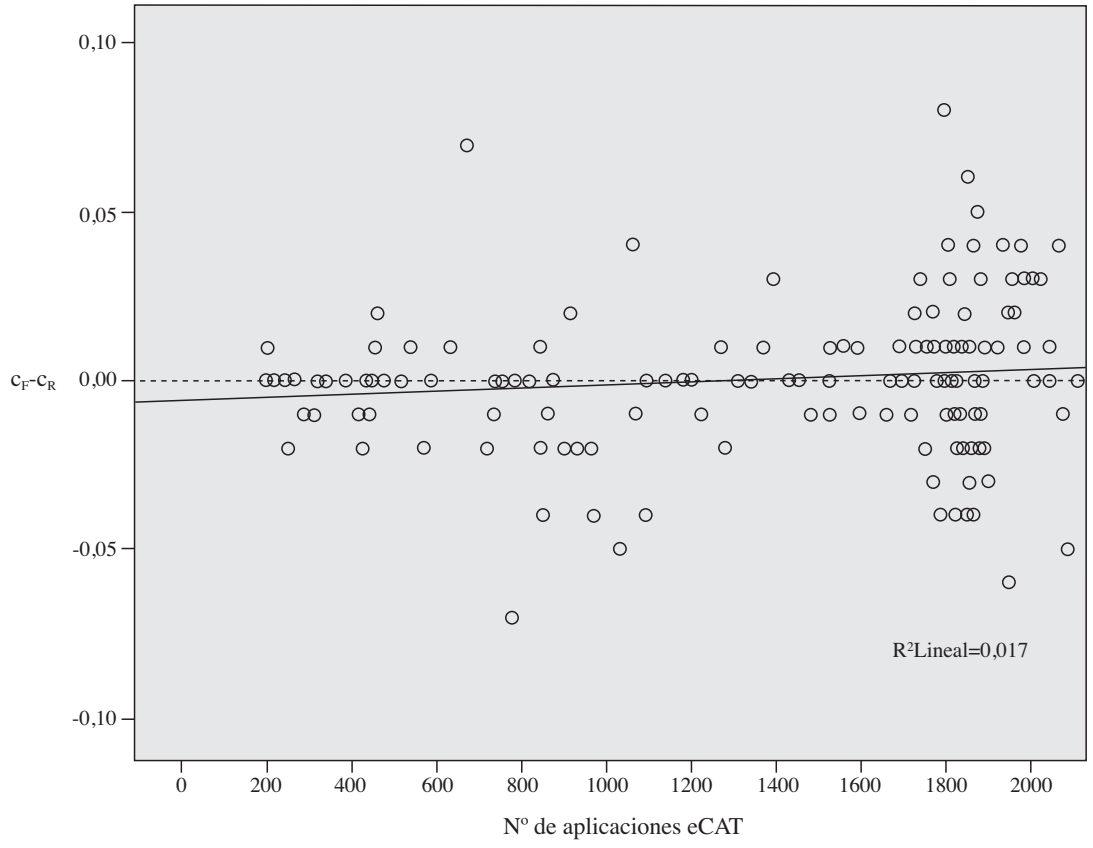

Figura 2. *Cambio relativo en a (cF-cR) como función del número de aplicaciones de eCAT*

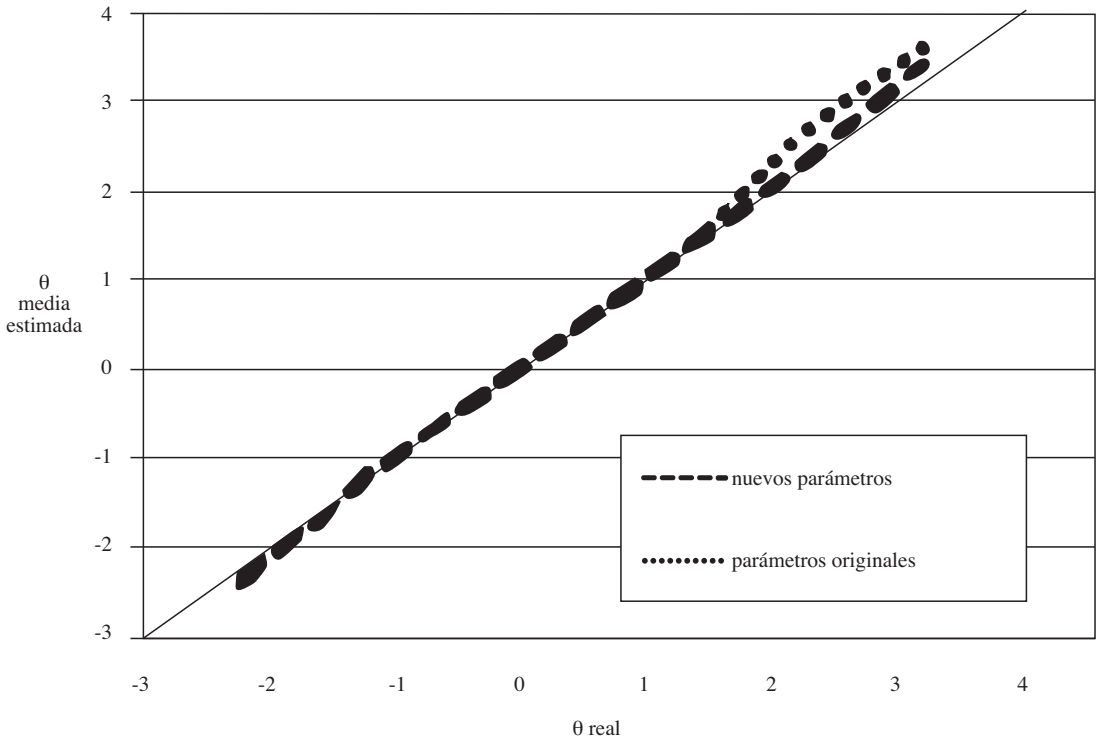

Figura 3. *: θ media estimada con ambos conjuntos de parámetros y θ real*

0,968, según se utilicen los nuevos parámetros o los originales). En la figura 3 se muestran los niveles medios estimados con ambos conjuntos de parámetros respecto de la θ real.

En ella se aprecia claramente que la repercusión del cambio de parámetros sobre la θ real es pequeña y que con los parámetros originales obtenidos en la aplicación de lápiz y papel se produce un sesgo positivo, sobreestimación del nivel de rasgo, en los niveles altos. También se obtuvo el gráfico de dispersión entre las θ estimadas con ambos conjuntos de parámetros (véase figura 4).

La correlación fue también muy alta en este caso (r= 0,993). Igualmente puede apreciarse que existen algunas diferencias para los evaluados con un nivel de habilidad alto (el nivel de habilidad estimado con los nuevos parámetros es algo menor).

Discusión y conclusiones

Los resultados presentados permiten apoyar la idea de que con el transcurso de las aplicaciones de eCAT los parámetros estimados en la aplicación original (formato lápiz y papel) pueden variar en la aplicación informatizada. La variación afecta a un conjunto importante de ítems y tiene una cierta repercusión en la estimación del nivel de rasgo de personas con alto nivel de inglés.

Más específicamente, el análisis de las diferencias en la estimación de parámetros en las fases pretest y *on-line* muestra que cuando el número de aplicaciones en eCAT es alto no se producen cambios sistemáticos en la estimación de los parámetros. Sin embargo sí aparecen cambios no sistemáticos, especialmente en los parámetros *a* y *c*. Además, el análisis FDI llevado a cabo ha mostrado que un número considerable de ítems (59) muestra cambios relevantes en sus parámetros. Se podría pensar que este cambio es debido a que, al presentar eCAT unas condiciones de control de las evaluaciones menor que en la muestra original, los ítems del banco se están difundiendo entre los evaluados, y por tanto los parámetros indican mayor facilidad de los ítems. Sin embargo la idea no parece apoyada por los datos, ya que las correlaciones entre el NCDIF (y las ventajas promedio) con las tasas de exposición de los ítems no son significativas.

Es difícil plantear razones específicas que puedan explicar los cambios no sistemáticos en los parámetros *a* y *c*. Creemos que las razones se pueden agrupar en dos tipos:

Razones relacionadas con cambios en el modo de aplicación: (a) En la aplicación original los ítems se aplicaban en formato de papel y lápiz y con orden prefijado, en grupos de aproximadamente 60 ítems con un rango variado de dificultad. En la aplicación adaptativa los ítems se presentan de forma informatizada y en grupos de 30 ítems ajustados en dificultad al nivel de la persona; (b) Las instrucciones sobre la penalización de los errores fue distinta en ambas aplicaciones: en eCAT nada se dice sobre ello, mientras que en la aplicación en formato de papel y lápiz se advirtió que los errores serían penalizados en la puntuación final; (c) En eCAT no se permite la revisión de respuestas.

Todas las diferencias anteriores entre aplicaciones pueden tener efectos en el modo de respuesta que interactúen con el nivel de habilidad y produzcan cambios en los parámetros *a* y *c* de los ítems. Esos cambios pueden depender de aspectos específicos del ítem (efectos de la posición que dependan de la posición del ítem en el test original, efectos de la posibilidad de revisión que dependan de lo necesario que resulte revisar el ítem, efectos de las instrucciones de penalización de los errores que pueden depender de los distractores específicos del ítem, etc.).

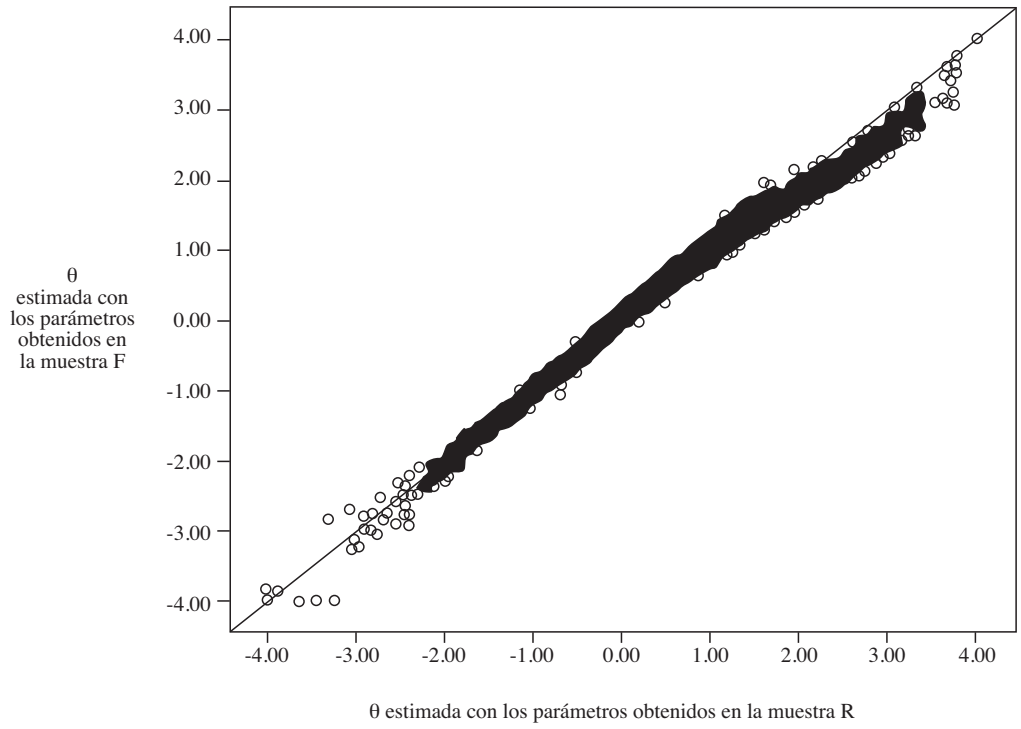

Figura 4. *: Gráfico de dispersión de las θ estimadas según los parámetros que se utilicen para la estimación (muestra R o F)*

Razones relacionadas con las diferencias en las características de los sujetos de las muestras: (a) La distribución de los niveles de inglés de la muestra de estudiantes chilenos puede ser distinta a la distribución de las personas que han respondido a eCAT en los diferentes contextos (selección de personal y evaluación del inglés en ámbitos universitarios). Los sesgos en la estimación de parámetros pueden ser distintos en función de la distribución del rasgo en las muestras de calibración; (b) También la motivación con la que ambas muestras responden a los tests puede ser distinta, dado que eCAT se aplica en procesos de selección de personal.

Independientemente de las razones, se ha comprobado también que el cambio en los parámetros tiene un ligero impacto sobre los niveles de rasgo estimados para los evaluados. Este impacto se centra fundamentalmente en la estimación de los niveles altos de rasgo y muestra cómo con los parámetros originales se produce una sobreestimación del nivel del evaluado. La razón de que el sesgo esté presente en los niveles altos puede deberse probablemente a que es mayor la proporción de ítems afectados entre los ítems de mayor dificultad y a que los cambios en los ítems aplicados a los participantes de alto nivel de habilidad son más sistemáticos.

Por todo ello, observando los cambios en los parámetros *a* y *c*, la presencia de ítems con FDI y la repercusión leve de los cambios en los niveles altos del rasgo, parece necesaria la actualización de los parámetros de los ítems. En este momento eCAT incorpora los parámetros estimados con la muestra total de personas (10462). Con ello, se han actualizado los parámetros del banco de ítems de forma que el peso de la muestra de eCAT en los valores de los nuevos parámetros sea dependiente del número de aplicaciones.

Adicionalmente a los resultados mostrados, el trabajo presentado muestra con claridad la necesidad de realizar estudios aplicados para comprobar el deterioro de parámetros de los TAIs en funcionamiento, y quizás en otros tipos de tests. El procedimiento seguido y las herramientas a aplicar para desarrollar un estudio de estas características puede servir a quiénes tienen la responsabilidad de que las estimaciones de rasgo sean lo más precisas posible.

Referencias

Abad, F.J., Olea, J., Ponsoda, V., Ximénez, C., y Mazuela, P. (2005). Efecto de las omisiones en la calibración de un test adaptativo informatizado. *Metodología de las Ciencias del Comportamiento, Supl*., 1-6.

Bolt, D. (2002). *Studying the potential of nuisance dimensions using bundle DIF and multidimensional IRT analyses*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans: LA.

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.

Donoghue, J.R., y Isham, S.P. (1998). A comparison of procedures to detect ítem parameter drift. *Applied Psychological Measurement, 22*, 33-51.

Glas, C.A.W. (2000). Item calibration and parameter drift. En W.J. van der Linden y C.A.W. Glas (Eds.): *Computerized adaptive testing: Theory and practice* (pp. 183-199). Boston, MA: Kluwer Academic Publishers.

Guo, F., y Wang, L. (2003). *Online calibration and scale stability of a CAT program*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago: IL.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144-149.

Hanson, B.A., y Beguin, A.A. (2002). Obtaining a common scale for IRT item parameters using separate versus oncurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*, 3-24.

Harmes, J.C., Parshall, C.G., y Kromrey, J.D. (2003). *Recalibration of IRT item parameters in a CAT: sparse data matrices and missing data treatments*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago: IL.

Holland, P.W. y Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. En H. Wainer y H.I. Braun (Eds.): *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

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

Mislevy, R.J., y Bock, R.D. (1990). BILOG *(version 3) [Computer software]*. Chicago: Scientific Software International.

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

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

Olea, J., y Ponsoda, V. (1996). Tests adaptativos informatizados. En J. Muñiz (Coord.): *Psicometría* (pp. 730-783). Madrid: Universitas.

Pommerich, M,. y Segall, D.O. (2003). *Calibrating CAT pools and online pretest items using marginal maximum likelihood methods*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago IL.

Ponsoda, V., Hontangas, P., Olea, J., Revuelta, J., Abad, F.J., y Ximénez, C. (2004). Los tests adaptativos informatizados: investigación actual. *Metodología de las ciencias del comportamiento, Supl*., 507-510.

Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 492-502.

Raju, N.S. (1999). DFIT5P: *A Fortran program for calculating dichotomous DIF/DTF [Computer software]*. Chicago, IL: Illinois Institute of Technology.

Rubio, V.J., 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.

Stocking, M.L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika, 55*, 461-475.

Thissen, D., Chen, W-H, y Bock, R.D. (2003). *Multilog (version 7) [Computer software]*. Lincolnwood, IL: Scientific Software International.

Wang, T., y Kolen, M.J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria, and an example. *Journal of Educational Measurement, 38*, 19-49.

Wells, C.S., Subkoviak, M.J., y Serlin, R.C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*, 77-87.

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* (pp. 221-244). Boston, MA: Kluwer Academic Publishers.

Zwick, R., y Thayer, D.T. (2002). Application of an empirical Bayes enhancement of Mantel-Haenszel differential item functioning analysis to a computerized adaptive test. *Applied Psychological Measurement, 26*, 57-76.