Actualización de los parámetros de los ítems de eCAT y propuesta de ampliación del banco

Francisco Abad

Julio Olea

Universidad Autónoma de Madrid

Informe de investigación eCAT 08-01

Diciembre de 2008

Cátedra UAM – IIC "Modelos y Aplicaciones Psicométricas"

Tabla de contenido

Resumen

El presente trabajo describe los principales trabajos sobre actualización de los parámetros de un banco de ítems. Tiene tres objetivos específicos en relación al test adaptivo informatizado de inglés eCAT (Olea, Abad, Ponsoda y Ximénez, 2004): 1) Evaluar el deterioro de los parámetros de su banco de ítems, 2) decidir qué parámetros de los ítems se mantienen y cuáles han de ser los nuevos valores de los que deban cambiar, y 3) evaluar la eficacia del diseño de calibración para los nuevos ítems.

Para determinar si ha habido o no deterioro de los parámetros de los ítems se ha estudiado su funcionamiento diferencial, resultando que 59 ítems de los 159 analizados mostraron funcionamiento diferencial entre la calibración y su aplicación en eCAT. Se ha estudiado la relación entre el cambio de las estimaciones de los parámetros y la tasa de exposición de los ítems y no se ha encontrado una relación apreciable, por lo que no parece que sea la exposición de los ítems lo que ha causado el cambio de valores de los parámetros. Se ha comprobado en un estudio de simulación el impacto en la zeta estimada de utilizar los parámetros antiguos o los nuevos y se ha comprobado que el impacto es de escasa relevancia en tamaño y afecta solo a las zetas superiores a dos.

En cuanto al segundo objetivo, se propone cambiar los parámetros de los 197 ítems del banco y proponer como nuevos valores los resultantes de la calibración concurrente de los ítems en ambas aplicaciones (calibración y eCAT). Se ofrece (Tabla 21) la relación de los nuevos parámetros de los 197 ítems de banco.

En cuanto al tercer objetivo, se propone un diseño de calibración para los nuevos ítems que se vayan a añadir. Se ha comprobado, mediante un estudio de simulación, que el diseño propuesto recupera razonablemente bien los parámetros reales de los ítems.

Abstract

This work describes the main available procedures to update item parameters. In relation to the computerized adaptive test of English eCAT (Olea, Abad, Ponsoda y Ximénez, 2004), its main aims are three: 1) To assess the parameter drift of its items, 2) to discover which item parameters should be kept and what should be updated, and 3) to assess the efficiency of the online calibration design suggested for the addition of new items to the item bank.

In order to explore whether or not item drift was present, an item differential functioning study was conducted. A total of 59 items out of 159 showed differential functioning between the calibration and the eCAT application. Parameter drift was found not to be related to item exposure rates. So, they should not be blamed for the parameter drift witnessed. Despite the high number of items flagged, a simulation study showed a small impact on ability estimates. In fact, ability estimates generated by the old parameter set hardly differed from those provided by the new set. Differences were small and affected only to abilities above two.

Concerning the second aim, the update of the whole 197 parameter set is advised. The new estimates were obtained by concurrent calibration of the calibration and operational samples. The new set of estimates is offered (in Table 21).

Concerning the third aim, an online calibration design is proposed in order to have ready new items to replace old ones if necessary. A simulation study was conducted to check the efficiency of the proposed design to recover the pretest item parameters.

1. DISEÑO INICIAL DEL BANCO

Para la elaboración del banco de ítems en que se sustenta eCAT se decidió redactar alrededor de 600 ítems de elección múltiple con 4 opciones de respuesta, y garantizar en lo posible la validez de contenido del banco, reflejando los principales aspectos que representan el conocimiento gramatical del idioma. Se establecieron las 7 categorías gramaticales y las 46 subcategorías que se muestran en la **Tabla 1** (más detalles en Olea, Abad y Ponsoda, 2002). Se elaboraron un total de 635 ítems, distribuidos como se indica en la **Tabla 1**.

Tabla 1
Distribución de ítems según categorías y sub-categorías gramaticales.

CATEGORÍA	SUBCATEGORÍAS.	Nº (%) DE ÍTEMS
Aspectos formales	2	17 (2.6)
Morfología	17	222 (34.9)
Morfosintaxis	1	7 (1.1)
Pragmática	2	20 (3.1)
Léxico	7	177 (27.8)
Sintaxis	14	82 (12.9)
Categorías compuestas	3	110 (17.3)
	46	635

Se decidió construir 15 subtests de 61 ítems cada uno (los máximos que podrían aplicarse en una hora lectiva en niveles bajos de inglés), 41 propios del subtest y 20 ítems de anclaje comunes a los diversos subtests. Tanto unos (ítems del test de anclaje) como otros (ítems propios del subtest) debían ser representativos del banco completo en una doble dirección: a) representar las categorías gramaticales proporcionalmente a su presencia en el banco completo, y b) incluir ítems de nivel heterogéneo de dificultad.

¿Cómo se eligieron los ítems de anclaje y los ítems propios de cada subtest?. Tanto en uno como en otro, la presencia proporcional de cada una de las categorías gramaticales se realizó como se indica en la **Tabla 2**. Dada la escasez de ítems en la categoría de Morfosintaxis, se decidió no representar dicha categoría en el test de anclaje.

Tabla 2
Representación de categorías gramaticales en el test de anclaje y en los ítems propios del subtest.

BANCO DE ÍTEMS			Nº DE ÍTEMS DEL SUBTEST "K"	
CATEGORÍA	**SUB**	**Nº (%) ÍTEMS**	**ANCLAJE**	**PROPIOS (Aprox.)**
Aspectos formales	2	17 (2.6)	1	1-2
Morfología	17	222 (34.9)	7	13-14
Morfosintaxis	1	7 (1.1)	0	0-1
Pragmática	2	20 (3.1)	1	1-2
Léxico	7	177 (27.8)	5	10-12
Sintaxis	14	82 (12.9)	3	5-6
Categorías compuestas	3	110 (17.3)	3	7-8
TOTAL	46	635	20	41

Para reflejar en el test de anclaje el diferente nivel de dificultad del banco se procedió de la siguiente forma:

a) Clasificar los ítems del banco en 10 categorías (deciles) de dificultad a partir de las valoraciones de la dificultad de cada ítem producidas por cinco jueces.
b) En la categoría gramatical menos frecuente en el banco (Aspectos Formales), elegir un ítem del nivel de dificultad más frecuente. Entre los ítems disponibles, se seleccionó aquel donde la fiabilidad interjueces sobre su dificultad fue mayor.
c) Hacer lo mismo con la siguiente categoría menos frecuente (Pragmática).
d) Seguir el proceso de tal forma que se cumplan dos criterios: elegir el número de ítems fijado para cada categoría y dos ítems de cada decil de dificultad.

En la **Tabla 3** aparecen sombreadas las casillas que reflejan la categoría gramatical y el nivel de dificultad de los 20 ítems de anclaje, seleccionados como tales según el procedimiento descrito.

Tabla 3
Selección de los ítems del test de anclaje.

Decil de Dificult.	**A-FOR (1)**	**MORF (7)**	**PRAG (1)**	**LEX (5)**	**SINT (3)**	**C-C (3)**
1 (fácil)	5 (1 ítem)	27	5 (1 ítem)	10	5	10
2	2	33 (2 ítems)	3	10	3	5
3	0	27 (1 ítem)	3	14	7	15 (1 ítem)
4	1	23 (1 ítem)	3	15 (1 ítem)	10	12
5	2	33 (2 ítems)	0	13	8	11
6	1	24 (1 ítem)	1	13	11 (1 ítem)	9
7	1	20	1	19 (1 ítem)	12 (1 ítem)	12
8	2	14	1	17	14 (1 ítem)	13 (1 ítem)
9	2	15	0	26 (1 ítem)	9	14 (1 ítem)
10 (difícil)	1	6	3	40 (2 ítems)	3	9
N	17	222	20	177	82	110

El mismo procedimiento se siguió para determinar los 41 ítems propios de cada subtest. De esta forma, como resultado final de este proceso disponemos de 15 subtests, cada uno con 20 ítems de anclaje comunes y 41 propios, equilibrados inicialmente en dificultad y representatividad de las categorías gramaticales. En el **Anexo 1** pueden consultarse la distribución conjunta de dificultad esperada y categoría gramatical para el test de anclaje y para cada uno de los subtests (un ítem del test de anclaje se modificó tras la aplicación del test en una muestra piloto). Más detalles pueden consultarse en Olea et al. (2002).

2. CALIBRACIÓN INICIAL DEL BANCO DE ÍTEMS

Los 225 ítems de cinco subtests fueron aplicados a una muestra total de 3224 estudiantes de la Universidad Católica de Chile (665, 660, 645, 636 y 618 estudiantes, respectivamente, respondieron a cada subtest). Esto significa que dispusimos de las respuestas de la muestra completa a los 20 ítems de anclaje y de los correspondientes tamaños muestrales a los 41 ítems propios de cada subtest. Los subtests se aplicaron en soporte de papel y lápiz, dando un tiempo global de 60 minutos para completar la prueba.

A partir de diversos análisis psicométricos se decidió eliminar los ítems del banco:

- que tenían bajas correlaciones biseriales con el total.
- en los que una opción incorrecta correlacionaba positivamente con el total del subtest.
- desajustados al modelo logístico de 3 parámetros (χ^2 con $p<0.01$).

Después de este proceso de depuración quedaron 197 ítems para formar definitivamente el banco. Sus parámetros fueron estimados 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). La distribución a priori *inicial* para los parámetros *a* era log-normal (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 estimaban 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 = 2.000) y beta para el parámetro *c* (media = 0.207, desviación típica = 0.023). Los parámetros estimados para cada uno de los ítems son los que se recogen en la **Tabla 4**. Más detalles pueden consultarse en Olea, Abad, Ponsoda y Ximénez (2004).

Tabla 4
Parámetros de los ítems de eCAT (D = 1.7).

Id	*a*	*b*	*c*
1	2.19	0.49	0.26
2	0.78	-1.96	0.21
3	1.64	0.58	0.21
8	1.45	-0.40	0.20
9	1.37	-0.84	0.21
10	1.53	0.51	0.19
11	1.79	-0.22	0.21
12	1.06	-1.13	0.21
15	1.17	-0.27	0.21
18	1.85	0.39	0.18
20	2.09	0.02	0.18
21	1.25	-0.62	0.21
22	1.24	0.00	0.20
23	1.05	0.60	0.20
24	1.06	1.02	0.25
29	0.90	0.13	0.22
30	0.91	0.71	0.22
33	2.03	0.75	0.13
35	0.94	-0.14	0.23
41	1.62	0.48	0.20
42	1.08	-0.89	0.21
48	1.16	0.48	0.21
50	1.70	-0.56	0.20
53	1.52	-0.41	0.21
57	1.21	0.15	0.21
58	1.58	0.22	0.21
60	1.40	0.78	0.18
61	1.60	0.11	0.21
62	1.39	1.07	0.19
63	1.09	0.35	0.23
66	1.77	0.81	0.24
67	1.63	0.71	0.22
69	1.22	1.28	0.24
73	0.93	0.18	0.23
74	1.32	2.28	0.21
76	1.02	1.24	0.29
79	1.53	0.96	0.18
80	1.03	0.27	0.21
84	2.03	0.32	0.16
85	1.45	0.92	0.20
86	2.13	0.14	0.19
88	1.33	-1.05	0.21
89	1.06	-0.22	0.23
95	0.98	-0.77	0.21
99	0.83	-2.71	0.21
100	0.61	-0.32	0.23
102	1.07	-0.09	0.20
104	1.50	0.86	0.17
105	1.47	-0.65	0.21
107	1.28	0.63	0.18
109	1.61	0.39	0.19
111	1.18	-0.44	0.23
112	1.21	0.34	0.26
115	0.70	0.16	0.24
116	1.37	-0.64	0.20
120	2.00	0.14	0.18
121	1.50	-0.07	0.21
123	1.29	-0.50	0.21
124	1.07	-0.80	0.21
125	1.27	-0.34	0.21
132	1.46	0.28	0.20
135	1.40	-0.07	0.19
145	0.95	-0.34	0.22
148	1.62	0.02	0.19
153	1.42	0.80	0.17
154	0.99	0.74	0.20
155	1.37	0.19	0.25
157	1.62	2.44	0.18
161	1.29	1.19	0.16
164	1.38	2.24	0.20
167	1.87	2.04	0.13
169	1.49	0.79	0.20
171	0.87	-1.09	0.21
176	1.17	1.47	0.20
179	1.58	0.51	0.22
181	1.17	-0.31	0.21
184	1.75	0.60	0.18
188	1.40	-0.46	0.20
189	1.08	0.65	0.20
190	1.17	1.25	0.19
192	0.68	-0.40	0.22
196	1.28	-0.02	0.23
197	1.28	-0.11	0.21
198	1.59	1.24	0.15
201	1.14	1.94	0.21
207	0.78	-0.36	0.23
209	1.47	-0.39	0.21
220	1.69	1.19	0.20
224	1.98	0.22	0.20
226	1.99	0.60	0.17
231	1.67	-0.21	0.21
232	1.36	-1.22	0.20
233	1.01	-0.20	0.21
237	1.31	-0.54	0.21
238	1.05	-1.11	0.21
240	1.00	-0.97	0.21
243	1.36	0.63	0.22
247	1.23	0.31	0.21
248	1.29	0.61	0.21
251	1.20	0.39	0.22
252	1.42	-0.50	0.21
253	1.08	1.78	0.25
257	1.58	0.69	0.23
258	0.87	-0.59	0.22
274	1.07	2.24	0.22
276	1.03	0.20	0.22
277	1.24	2.31	0.23
278	1.22	2.91	0.22
280	1.38	1.58	0.25
281	1.00	1.05	0.22
285	1.25	1.07	0.20
296	1.32	1.22	0.18
297	1.49	-0.10	0.19
300	1.09	0.62	0.20
307	2.20	-0.03	0.23
308	1.20	2.00	0.25
313	1.45	-1.26	0.21
326	1.06	-0.18	0.21
330	1.61	0.40	0.23
339	0.97	-0.80	0.23
343	1.07	-1.56	0.21
344	1.04	-0.11	0.21
348	1.30	-0.61	0.21
350	1.05	1.17	0.26
358	1.29	2.31	0.28
359	1.17	0.34	0.22
363	1.14	1.09	0.20
364	0.89	0.69	0.21
367	0.82	0.41	0.25
369	1.90	0.77	0.16
372	1.77	-0.57	0.20
374	1.73	0.83	0.20
375	0.93	0.87	0.23
377	1.23	1.14	0.19
381	1.88	0.26	0.19
384	1.37	-0.02	0.22
390	1.67	-0.42	0.21
391	1.26	-1.61	0.21
393	1.91	0.43	0.18
394	1.17	0.60	0.20
403	1.45	0.01	0.20
406	1.48	0.39	0.20
412	1.30	-0.43	0.20
415	0.94	-2.45	0.21
416	1.08	0.38	0.22
421	1.64	-1.27	0.21
422	0.97	-0.98	0.21
424	1.10	-1.05	0.21
427	1.06	0.19	0.22
430	1.14	-0.69	0.21
433	1.12	0.75	0.23
438	1.02	-0.11	0.22
442	1.46	1.24	0.19
447	1.25	0.25	0.28
454	1.35	-0.39	0.22
462	1.55	0.00	0.20
479	1.03	1.26	0.22
480	1.13	-0.39	0.21
494	1.75	0.92	0.17
498	1.55	1.20	0.14
505	0.82	-0.48	0.21
518	1.07	2.26	0.21
519	1.01	-0.76	0.21
524	1.38	2.10	0.15
535	1.71	0.11	0.22
536	1.34	0.22	0.23
545	1.22	3.42	0.19
550	1.50	0.36	0.20
554	1.40	2.16	0.21
567	0.91	-0.11	0.23
571	1.24	2.04	0.17
584	1.11	2.03	0.23
590	1.32	-0.63	0.21
591	1.32	-0.68	0.21
593	0.90	0.49	0.24
595	1.37	-0.89	0.20
596	1.25	-0.38	0.21
597	1.01	-0.50	0.21
598	1.49	-0.37	0.20
599	1.57	-0.80	0.20
606	0.43	-1.72	0.21
607	1.04	0.80	0.11
612	0.75	-0.13	0.22
615	1.40	0.25	0.20
617	1.41	-0.21	0.21
621	1.27	1.24	0.20
623	1.14	-1.00	0.22
625	1.34	-1.40	0.21
628	1.30	2.94	0.11
635	1.26	0.12	0.19
636	1.68	-0.73	0.20
641	1.20	-1.38	0.22
644	0.94	-0.57	0.21
653	1.57	0.33	0.22
661	0.92	-0.59	0.22
662	1.02	-0.50	0.20
663	1.37	0.01	0.21

3. ESTUDIO DEL DETERIORO DE LOS PARÁMETROS

Como indicamos recientemente en una revisión sobre el estado actual de la investigación sobre TAIs, el estudio de procedimientos eficientes para el mantenimiento y la renovación de los bancos de ítems es uno de los desafíos principales para que este tipo de pruebas sigan aplicándose (Ponsoda, Hontangas, Olea, Revuelta, Abad y Ximénez, 2004). Cabe pensar que, después de las sucesivas aplicaciones del banco mediante los algoritmos adaptativos incorporados en eCAT, algunos parámetros de los ítems han podido cambiar, y por tanto es necesaria su actualización. La expresión en inglés para referirnos a estos cambios es la de *Item Parameter Drift*. Es un problema bastante estudiado en TRI que, a pesar de la teórica propiedad de invarianza de los estimadores, los parámetros de los ítems varían en sucesivas aplicaciones del mismo test o del mismo TAI (v.gr., Donoghue y Isham, 1998; Glas, 2000; Wang y Kolen, 2001). En nuestro caso concreto, las diferencias pueden ser debidas a varias razones:

Diferencias en las condiciones de aplicación:

1. Los ítems iniciales se presentaron en formato de papel y lápiz y con un tiempo disponible diferente que en el formato informatizado.
2. 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.
3. En eCAT no se permite la revisión de respuestas.
4. En la primera aplicación se muestrearon bien las diferentes categorías de contenido de los ítems para construir los diferentes subtests, cosa que no se hace en la aplicación adaptativa.
5. En la muestra de calibración todos los sujetos respondieron a todos los ítems de un subtest, ordenados en dificultad previsible, mientras que en eCAT los ítems aplicados se ajustan en dificultad al nivel de habilidad que van manifestando los evaluados.

Diferencias en las características de los sujetos de las muestras:

6. 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.
7. 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.
8. Es posible que, en algunos de los contextos de aplicación de eCAT se hayan difundido algunos de los ítems, algo que sabemos seguro que no ocurrió en la muestra chilena.

Una conjunción de estos factores puede llevar a cambios significativos en las estimaciones de los parámetros de dificultad, discriminación y pseudoazar. En tests fijos se han propuesto diversos procedimientos para estudiar la variación de los parámetros de los ítems. Por ejemplo, existen procedimientos basados en χ^2 para comprobar de forma simultánea si los 3 parámetros de un ítem son estadísticamente equivalentes en dos aplicaciones distintas (Lord, 1980); otros procedimientos obtienen mediante integración el área delimitada por las funciones de respuesta obtenidas para un ítem en dos calibraciones. También se han propuesto métodos que comparan los valores χ^2 de cada calibración por separado con los obtenidos en la calibración conjunta (Donoghue y Isham, 1998). Otros aplican procedimientos de detección del Funcionamiento Diferencial del Ítem (FDI) derivados, por ejemplo, del método de Mantel-Haenzel (Holland y Thayer, 1988).

El estudio del cambio de parámetros de los ítems cuando se aplican en un TAI consiste en 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*) o cuando se obtiene a partir de las respuestas dadas en las sucesivas aplicaciones del TAI (etapa *on- line*). El problema de la variación de las estimaciones de los parámetros de los ítems es especialmente relevante en la aplicación de TAIs, ya que las respuestas de los sujetos tienen más relevancia en la estimación de su nivel θ que en un test fijo (usualmente se presenta un menor número de ítems con elevado nivel de información).

Cualquier variación en los parámetros de los ítems tiene importantes consecuencias en la métrica en la que se mide θ.

La presentación adaptativa de ítems tiene algunos problemas particulares en el estudio de los cambios de parámetros. 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. 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). En este sentido, es importante el estudio de Guo y Wang (2003) que propone estudiar el impacto del deterioro de los parámetros en las puntuaciones de los evaluados mediante un estudio de simulación.

En el presente informe se valorará el cambio de parámetros de los ítems en las actuales aplicaciones de eCAT. Para ello se estudiará el funcionamiento diferencial de estos mediante una estrategia de TRI y la repercusión del FDI en las estimaciones obtenidas del nivel de rasgo.

4. RECALIBRACIÓN DEL BANCO COMPLETO

A medida que se aplica el TAI en numerosas ocasiones, disponemos de nueva información para recalibrar los ítems, de tal manera que las estimaciones de sus parámetros sean más precisas e incorporen las variaciones esperables por el nuevo modo de aplicación adaptativa (p.e., tras el cambio de formato de lápiz y papel a informatizado).

Sin embargo, no resulta sencillo emplear la nueva información en la nueva estimación de parámetros. Como hemos dicho, la nueva información es incompleta dado que el banco incluye más ítems que el test y que se aplican un porcentaje elevado de ítems distintos a los evaluados. Además, dada la naturaleza adaptativa de estos tests, existe una importa 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 es disponer de una muestra amplia con distribución de rasgo uniforme (Stocking, 1990). 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 se establecen valores iniciales inadecuados para los parámetros (Pommerich y Segall, 2003).

Algunos autores proponen tratar el problema de la estimación como un problema general de análisis de matrices con datos incompletos. En muchas circunstancias disponemos de matrices incompletas de respuestas a los ítems, cuando deben estimarse los parámetros de dichos ítems y los niveles θ de los evaluados. Es el caso de los datos disponibles tras la aplicación de un TAI, los obtenidos a partir de diseños de anclaje, donde no todos los sujetos responden a todos los ítems, o los que se derivan de condiciones de aplicación donde se producen muchas omisiones, por ejemplo porque se penalizan fuertemente los errores.

En la aplicación de los TAIs, las omisiones representan un porcentaje elevado de los datos disponibles para la calibración de los ítems (normalmente por encima del 80 % de los datos disponibles). Estas condiciones hacen que sea muy complicada la calibración de los ítems si no se aplica algún método de imputación a los datos perdidos. No son muchos los trabajos que se han realizado sobre los métodos de estimación de los parámetros de los ítems más apropiados en condiciones de tasas elevadas de omisiones. Harmes, Parshall y Kromrey (2003) propusieron un procedimiento para estimar los

parámetros de los ítems a partir de las respuestas a un TAI. Desde un punto de vista estadístico, consideraron la situación como un problema de datos perdidos de forma no aleatoria y lo intentaron resolver mediante técnicas de imputación múltiple para completar los datos.

En el presente informe se estudiará la posibilidad de calibrar el banco a través del programa MULTILOG (Thissen, Chen y Bock, 2002) mediante procedimientos de calibración concurrente (más sencillos). Dados los resultados de estudios previos obtenidos con BILOG, se atenderá a:

1. Los posibles problemas de convergencia.
2. Las transformaciones métricas (equiparación) necesarias para que la escala de los parámetros en la recalibración sea la misma que la de los parámetros originales.
3. La calidad de la recuperación de parámetros en función del número de aplicaciones (mediante un estudio de simulación).

5. ACTUALIZACIÓN DEL BANCO: CALIBRACIÓN DE NUEVOS ITEMS (ítems pretest)

Para que las propiedades psicométricas de las estimaciones que se realizan con un TAI se mantengan es necesario que sus ítems sean periódicamente revisados y sustituidos. Un ítem puede retirarse si en nuevas muestras sus propiedades psicométricas se han modificado (*drift parameter*) o, con objeto de mantener la seguridad del banco, tras un número predefinido de aplicaciones.

El proceso inevitable de reponer el banco añadiendo nuevos ítems es costoso, pues se requieren nuevas muestras para estimar sus parámetros. Se han propuesto diversos métodos, que no son necesariamente excluyentes, para reducir los costes de dicha calibración. Un grupo de trabajos pretenden obtener estimaciones precisas de los nuevos ítems con muestras reducidas y procedimientos más económicos; por ejemplo, emplear los juicios de expertos sobre la dificultad junto con las respuestas de los sujetos (Swaminathan, Hambleton, Sireci, Xing y Rizavi, 2003), estimar los parámetros a partir de indicadores psicométricos clásicos (Huang, Kalohn, Lin y Spray, 2000) o establecer diseños muestrales óptimos para seleccionar a los sujetos de nivel de habilidad apropiado para la estimación de los parámetros de los ítems (Brumfield, Burroughs y Luecht, 2001). Un problema de esta última perspectiva es que una vez en el TAI, si los ítems se aplican adaptativamente, los parámetros de los ítems son más difíciles de actualizar incluso si se han aplicado a un conjunto numeroso de personas (Do, Chuah y Drasgow, 2004).

Otra forma, parece que más eficiente, de actualizar el banco es presentar en cada aplicación nuevos ítems (ítems *pretest*) junto con los que ya se encuentran operativos en el banco. Es lo que se conoce como *calibración on-line*. El conocimiento de los parámetros de los ítems operativos permite estimar los parámetros de los ítems *pretest* incluso si los examinados han respondido a conjuntos de ítems muy distintos (que es lo que ocurre en una aplicación adaptativa). Asimismo, esta calibración *on-line* se puede utilizar para revisar o mejorar las estimaciones de los ítems operativos bajo la consideración de las nuevas respuestas.

Los procedimientos de calibración *on-line* que se han propuesto pueden clasificarse en varias categorías. Aquí revisaremos los procedimientos que se basan en el algoritmo EM. Existen otros procedimientos no basados en el algoritmo EM (p.e., de estimación

MCMC o procedimientos no paramétricos) pero su exposición en este trabajo va más allá de los objetivos de este estudio (más información puede encontrarse en Segall, 2003, y en Krass y Williams, 2003).

Entre los procedimientos basados en el algoritmo EM podemos realizar la siguiente distinción:

- Calibración on-line *manteniendo* los parámetros de los ítems operativos.
- Calibración on-line *actualizando* los parámetros de los ítems operativos.

5.1. Calibración on-line *manteniendo los parámetros de los ítems operativos.*

Muchos de los procedimientos de calibración *on-line* parten del procedimiento EM original. En general, el objetivo de estos procedimientos es mantener los parámetros de los ítems operativos. Recordemos que en ese procedimiento los parámetros de los ítems se estiman mediante un procedimiento iterativo. Cada ciclo *c* EM tiene 2 fases:

Fase E. Se estima la distribución posterior de la habilidad dadas las respuestas de cada sujeto *j* y los parámetros de todos los ítems estimados en el ciclo anterior *c-1*:

$$g(\theta \mid x_j, \beta_{c-1})$$

Fase M. se estiman los parámetros del ítem *i* que maximizan la función de verosimilitud en ese ítem; para ello, se asumen las distribuciones posteriores estimadas en el paso *E*:

$$L = \prod_{j=1}^{N} \int L(x_{ij} \mid \theta) g(\theta \mid x_j, \beta_{c-1}) d\theta$$

Los tres trabajos fundamentales que comparan procedimientos basados en el algoritmo *EM* son los de Ban, Hanson, Wang, Yi y Harris (2001), Ban, Hanson, Yi y Harris (2002) y Pommerich y Segall (2003). Estos procedimientos fueron los siguientes:

- *Método OEM* (propuesto por Wainer y Mislevy, 1990). Se utiliza el algoritmo EM (con un solo ciclo) para estimar los parámetros de los ítems. Primero, se calcula en un único paso *E* la distribución posterior de la habilidad. Esto se hace sólo a partir de los parámetros de los ítems operativos. Posteriormente, en el

paso *M* se calculan los parámetros de los ítems *pretest* que maximizan la verosimilitud de los datos.

- *Método MEM* (propuesto por Ban et al., 2001). Igual al método OEM, pero se permite más de un ciclo, pues para estimar la distribución posterior de la habilidad se utilizan también los parámetros estimados para los ítems *pretest.*

- *BILOG o BILOG-MG fijando los parámetros de los ítems operativos o imponiéndoles distribuciones previas muy fuertes***.** Esta aproximación es la más sencilla. Se utiliza un procedimiento de calibración concurrente (se calibran todos los ítems a la vez siguiendo el algoritmo EM original). Para los parámetros de los ítems operativos se fijan a sus valores conocidos o se les imponen distribuciones previas que restringen mucho sus valores posibles (lo que es equivalente a tratar esos parámetros como fijos en el proceso de estimación). Uno de los problemas de BILOG es que cuando se calibran datos de un TAI, bastante frecuentemente, no se alcanza la convergencia (Ban et al., 2001).

- *Método B de Stocking* (propuesto por Stocking, 1988). Cada examinado responde a ítems *operativos*, ítems *pretest* e ítems de *ancla* (ítems operativos que no se van a aplicar de forma adaptativa). Se ejecuta en dos pasos. En el primero, se calibran los parámetros de los ítems *pretest* junto con los de anclaje. Para ello, se estima el nivel de habilidad de los examinados a partir de las respuestas a los ítems operativos. A continuación, se fijan las habilidades estimadas como si fueran fijas y se estiman los parámetros de los ítems *pretest* y de *anclaje*. En el segundo paso, se transforman linealmente los parámetros de los ítems de tal manera que los parámetros estimados para los ítems de anclaje se aproximen a sus valores originales. Este procedimiento no parece teóricamente recomendable porque asume los niveles de rasgo estimado como verdaderos y además requiere un procedimiento de equiparación posterior.

En todos los casos, los procedimientos se comparan después de que los ítems *pretest* hayan sido aplicados a un número de sujetos. Para estudiar estos métodos, se utilizan procedimientos de simulación y se toman distintas medidas de precisión que indican la calidad en la recuperación de los parámetros de los ítems y también de sus CCI

(generalmente considerando el intervalo de rasgo en la muestra a la que se está aplicando el TAI). Ban, et al. (2001) trabajan con un TAI sin control de la exposición de los ítems y con la selección de ítems por el método de máxima información, concluyen que el método que funciona mejor es el MEM (y el que peor, el OEM) incluso aún cuando en el método de Stocking se aplicaban más ítems (los ítems de ancla) a los examinados. En Ban et al. (2002) se trabaja con un TAI en el que se seleccionaban los ítems en función de su parámetro *b* y se incluía control de la exposición de los ítems (en cada paso, se elegía aleatoriamente entre los 2 ítems candidatos y, adicionalmente, se imponían tasas máximas de exposición). Los autores llegaron a las mismas conclusiones. En ninguno de esos estudios se analiza el efecto en la precisión del número de ítems operativos que se incluyen en el TAI ni de la proporción de ítems pretest (o de anclaje) que incluye el test.

Recientemente, Kim (2006a) ha propuesto un nuevo procedimiento, extensión del método MEM, el MWU-MEM (Multiple Weights Updating – Multiple EM cycles). Como es sabido, la distribución posterior de θ depende de la distribución previa que se escoja para θ. En los procedimientos descritos hasta el momento se asume que θ se distribuye normalmente N (0,1). Esto puede ser un error porque al establecer esa distribución previa se fija doblemente la métrica de los parámetros (la métrica de los parámetros de los ítems pretest ya debería quedar fijada si se dan los parámetros de los ítems operativos como fijos). De hecho este es otro de los problemas de BILOG: asume una distribución previa para θ con media *0* y desviación típica *1*. Dicho de otra manera, cuando se fijan los parámetros de los ítems de anclaje la media y la desviación típica deberían estimarse a partir de los datos.

Justamente esto es lo que se hace en el nuevo procedimiento MWU-MEM. En este procedimiento, en cada paso M se estiman y actualizan la media y la desviación típica de la distribución previa de θ. Este procedimiento puede implementarse con el lenguaje ICL (Hanson, 2002) y no ha sido aplicado por el momento en el contexto de la actualización de parámetros en TAIs.

Por tanto, hay que ser cauto en el uso de BILOG o MULTILOG cuando se asumen parámetros fijos para los ítems operativos. Al asumir una distribución normal N (0,1) para el rasgo y simultáneamente fijar los parámetros de los ítems de *anclaje* a unos determinados valores concretos se introducen sesgos en la estimación de los parámetros.

5.2. Calibración on-line *actualizando los parámetros de los ítems operativos*

Algunos autores señalan el riesgo de fijar los parámetros de los ítems operativos. Además de los problemas comentados, una razón señalada (p.e., Partchev y Steyer, 2004) es que los ítems operativos aplicados son los que tienen mayor sesgo positivo en la estimación del parámetro *a* y esto podría producir sesgos en el establecimiento de la métrica.

Por ello, otra posibilidad es plantearse que los parámetros de algunos ítems operativos se actualicen. En el trabajo de Pommerich y Segall se realiza un estudio de simulación emulando el test adaptativo CAT-ASVAB. En esa prueba se incluye control de la exposición y el método de selección de ítems de máxima información; el banco se actualiza incluyendo un único ítem de prueba (pretest) dentro del CAT operativo. Se simulan examinados de distribuciones con distinto nivel medio de rasgo. Ellos distinguen entre tres tipos de ítems:

a.) *pretest*: ítems que se van a *calibrar por primera vez.*

b.) *operativos*: ítems *ya calibrados* pero que requieren revisión.

c.) *de anclaje:* ítems *ya calibrados* que no requieren revisión. Puesto que no se aplican frecuentemente, sus parámetros se comnsideran estables a través de aplicaciones sucesivas.

Los autores concluyen que la calibración MML mediante BILOG funciona bastante mal para algunos ítems (por ejemplo, hay ausencia de convergencia), probablemente, según ellos, por unos incorrectos valores iniciales y por los pequeños tamaños muestrales para algunos ítems.

Finalmente, un nuevo desafío para la investigación es cómo combinar el uso de un método óptimo de calibración (entre los anteriores) con una selección óptima de los ítems *pretest* a aplicar a cada examinado. Diversos estudios ya han mostrado que se puede mejorar la estimación de los parámetros de los ítems seleccionando a la muestra de examinados para la calibración de forma óptima (Holman y Berger, 2001; Stocking, 1990). Mientras que mucha investigación se ha realizado en la línea de seleccionar óptimamente los ítems para estimar los parámetros de los examinados (los TAIs), poco se ha hecho en la línea complementaria: *seleccionar óptimamente los examinados para estimar los parámetros de los ítems.* Alguna propuesta se ha hecho en esta línea. Buyske (1998) observa que cada examinado es *óptimo* para calibrar un conjunto de

ítems (generalmente aquellos cuyo nivel de dificultad está ligeramente por encima o ligeramente por debajo de su nivel de habilidad). Los ítems *pretest* pueden ser aplicados al examinado después de que responda al TAI operativo, cuando ya disponemos de una estimación precisa de su nivel de habilidad. Al principio de su proceso de calibración, los ítems *pretest* pueden ser seleccionados al azar. Una vez el ítem ha sido aplicado unas cuantas veces puede tenerse una idea inicial de su dificultad. A partir de entonces, para cada examinado, se podrá seleccionar entre los k ítems los que, por su nivel de habilidad, puedan ser mejor calibrados.

En el presente informe se estudiará la posibilidad de calibrar el banco utilizando el programa ICL (Hanson, 2002). Se propone un diseño de calibración asumiendo parámetros fijos para los ítems operativos (mediante el método MWU-MEM) y se estudia mediante un estudio de simulación, la eficacia previsible de este diseño en la recuperación de parámetros de los ítems pretest.

6. Objetivos del presente estudio

Nuestros objetivos en el presente estudio son los siguientes:

- *Evaluar el deterioro de los parámetros de eCAT.* Se valorará la necesidad de actualizar los parámetros de los ítems mediante el estudio del funcionamiento diferencial de estos. Se evaluará la repercusión del FDI de los ítems en las estimaciones obtenidas del nivel de rasgo.
- *Decidir que parámetros de los ítems se mantienen para eCAT.* Si no hay cambios, se mantendrán los parámetros de los ítems originales. Si se requiere actualización se utilizará un diseño de calibración concurrente donde todos los parámetros pueden cambiar. Caben dos posibilidades:
  - Si los cambios no son muy importantes: Se utilizarán los parámetros estimados con la muestra completa (Lapiz y papel y eCAT). Dados los resultados obtenidos en estudios previos, esta es la estrategia más segura puesto que para todos los ítems se dispondrá de un conjunto amplio de respuestas (las de la aplicación original de Chile).
  - Si los cambios parecen muy importantes: Se utilizarán, para aquellos ítems para los que haya *suficiente muestra,* los parámetros estimados exclusivamente con la muestra de eCAT. Puesto que en estudios previos se han encontrado dificultades en la recuperación de parámetros, se estudiará mediante un pequeño estudio de simulación, la eficacia de esta estrategia.
- *Evaluar la eficacia del diseño de calibración para los nuevos ítems.* Finalmente, se propone un diseño de calibración para los ítems pretest que se vayan a añadir. En concreto, se propone asumir parámetros fijos para los ítems operativos (mediante el método MWU-MEM y el programa ICL) y se estudia, mediante un pequeño estudio de simulación, la eficacia previsible de este diseño en la recuperación de parámetros de los ítems.

7. ANÁLISIS DE LAS APLICACIONES DE E-CAT

7.1. ANÁLISIS DESCRIPTIVOS SOBRE NIVELES DE INGLÉS, ERROR TÍPICO, TIEMPO EMPLEADO Y TASAS DE EXPOSICIÓN

Disponemos de la información de 7254 sujetos que en diversas aplicaciones han respondido a eCAT en una condición común a todos de 30 ítems como criterio de parada. En la **Tabla 5** se describen algunos resultados descriptivos de estas aplicaciones sobre el error típico, θ estimada y tiempo medio de aplicación (media a nivel de ítem). El error típico medio es de 0.22, equivalente a una fiabilidad de 0.96. 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 ha empleado 20 segundos por ítem, lo que lleva a un tiempo medio de aplicación de los ítems de 10 minutos

Tabla 5
Estadísticos descriptivos.

	N	*Mínimo*	*Máximo*	*Media*	*Desv. Típ.*
Zeta estimada	7254	-2.64	4.00	.67	.93
Error típico	7254	.16	.84	.22	.04
Tiempo medio por ítem	7254	1.00	37.83	20.28	5.44

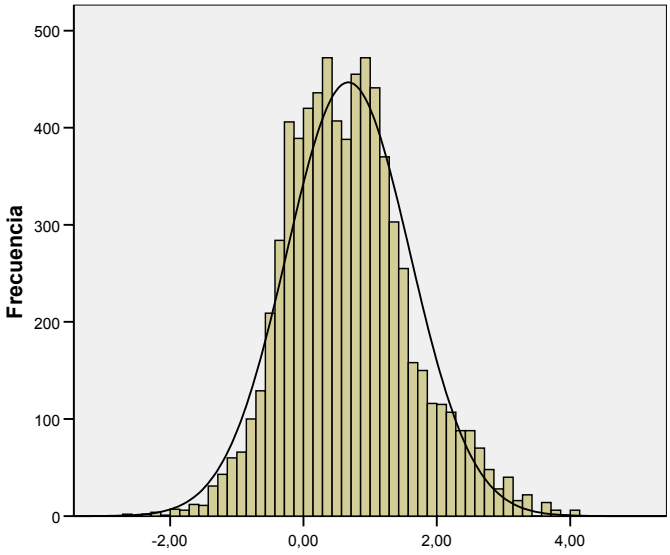
En la **Tabla 6** se recoge un baremo en centiles de las puntuaciones θ estimadas.

Tabla 6
Baremo en centiles.

θ	*Centil*
-1.30	1
-.41	10
-.11	20
.14	30
.37	40
.62	50
.86	60
1.08	70
1.36	80
1.89	90
3.08	99

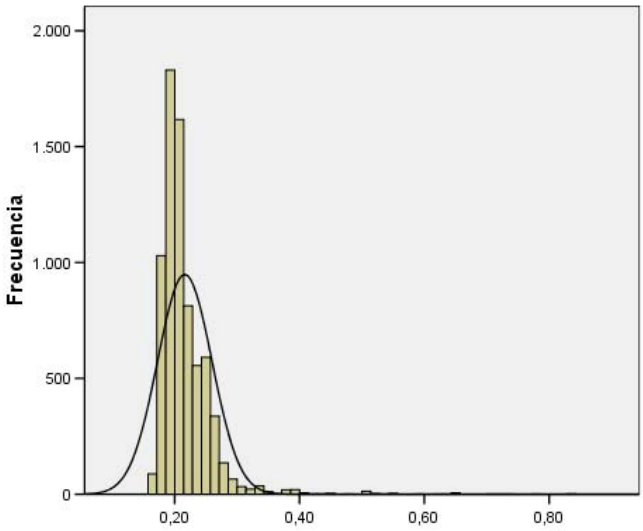
La distribución de puntuaciones, de errores típicos y de tiempo medio por ítem empleado se muestran en las **figuras 1, 2 y 3**:

Figura 1
Distribución de θ



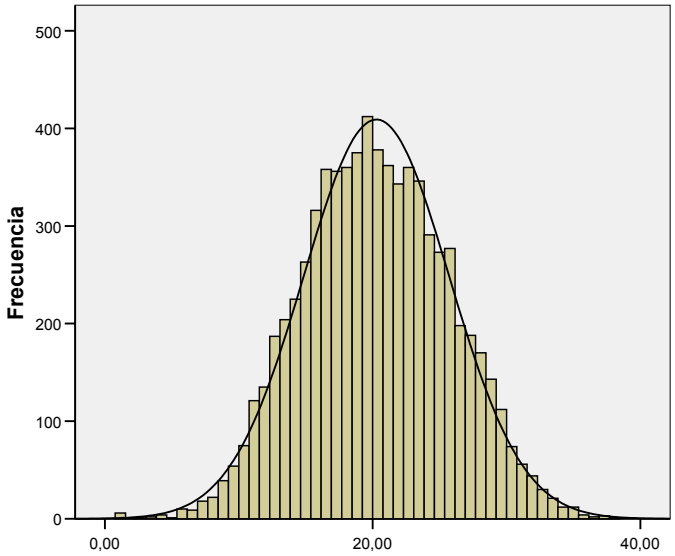
Zeta estimada

Figura 2
Error típico de estimación de θ



Error típico de medida

Figura 3
Distribución del Tiempo medio por ítem



La correlación entre los niveles θ estimados y el tiempo total empleado en el test fue de -0.166, una correlación inversa y significativa ($p < 0.01$), lo que indica que la gente de mayor nivel de habilidad, además de mostrar un mayor rendimiento, tiende a ser más rápida (aunque la relación es pequeña).

En la **Tabla 7** se muestra, por orden, la cantidad (y la proporción) de veces que se presentaron cada uno de los 180 ítems seleccionados en alguna ocasión. 17 ítems no fueron nunca seleccionados (ítems 15, 29, 35, 100, 102, 115, 145, 192, 207, 258, 339, 505, 567, 606, 612, 644 y 661). Son ítems con bajo parámetro de dificultad, no ajustados a los niveles de inglés de los evaluados que responden a eCAT, con niveles medios y altos en el manejo del idioma.

Tabla 7

Tasas de exposición (T.E.) y número de presentaciones (Pre) de los ítems de eCAT.

ID	*Pre*	*T.E.*	330	1817	0.25	12	1469	0.20	232	456	0.06
621	2096	0.29	24	1814	0.25	105	1447	0.20	422	449	0.06
280	2074	0.29	393	1813	0.25	663	1431	0.20	30	442	0.06
243	2059	0.28	1	1811	0.25	238	1421	0.20	197	435	0.06
161	2055	0.28	550	1811	0.25	571	1387	0.19	433	425	0.06
62	2031	0.28	381	1810	0.25	599	1360	0.19	438	419	0.06
296	2030	0.28	80	1809	0.25	248	1332	0.18	625	386	0.05
190	2009	0.28	344	1807	0.25	519	1306	0.18	233	339	0.05
60	1998	0.28	132	1804	0.25	554	1271	0.18	125	327	0.05
85	1993	0.27	23	1801	0.25	363	1267	0.17	247	325	0.04
167	1973	0.27	427	1800	0.25	384	1215	0.17	391	320	0.04
442	1971	0.27	18	1798	0.25	116	1196	0.16	623	312	0.04
153	1965	0.27	84	1793	0.25	201	1178	0.16	641	290	0.04
377	1947	0.27	403	1793	0.25	240	1134	0.16	424	270	0.04
79	1946	0.27	86	1791	0.25	164	1088	0.15	196	254	0.04
76	1937	0.27	607	1785	0.25	308	1085	0.15	415	254	0.04
169	1934	0.27	224	1784	0.25	454	1064	0.15	48	221	0.03
104	1920	0.26	120	1782	0.25	157	1054	0.15	95	221	0.03
285	1910	0.26	535	1776	0.24	253	1029	0.14	22	209	0.03
498	1910	0.26	297	1775	0.24	74	969	0.13	42	209	0.03
257	1885	0.26	121	1769	0.24	375	959	0.13	189	202	0.03
220	1880	0.26	462	1762	0.24	584	925	0.13	343	202	0.03
281	1876	0.26	524	1758	0.24	590	914	0.13	430	195	0.03
10	1872	0.26	148	1757	0.24	536	895	0.12	251	186	0.03
3	1869	0.26	176	1756	0.24	237	868	0.12	300	169	0.02
198	1867	0.26	20	1748	0.24	421	868	0.12	124	165	0.02
179	1865	0.26	307	1739	0.24	597	859	0.12	596	147	0.02
653	1865	0.26	326	1737	0.24	155	847	0.12	2	130	0.02
41	1858	0.26	107	1728	0.24	412	842	0.12	364	129	0.018
374	1858	0.26	231	1724	0.24	277	841	0.12	57	98	0.014
479	1854	0.26	11	1717	0.24	591	816	0.11	359	66	0.009
494	1853	0.26	135	1717	0.24	9	784	0.11	73	39	0.005
67	1849	0.25	390	1714	0.24	358	775	0.11	112	38	0.005
66	1846	0.25	276	1707	0.24	595	748	0.10	447	38	0.005
69	1842	0.25	372	1686	0.23	518	737	0.10	171	34	0.005
406	1841	0.25	598	1683	0.23	348	731	0.10	99	22	0.003
350	1839	0.25	53	1664	0.23	274	718	0.10	593	21	0.003
615	1839	0.25	617	1656	0.23	628	670	0.09	480	17	0.002
33	1838	0.25	50	1654	0.23	394	632	0.09	416	13	0.002
58	1833	0.25	636	1587	0.22	313	586	0.08	111	11	0.002
109	1830	0.25	8	1581	0.22	545	570	0.08	63	5	0.001
369	1828	0.25	209	1576	0.22	123	538	0.07	181	5	0.001
226	1826	0.25	188	1548	0.21	88	518	0.07	367	2	0.000
184	1825	0.25	89	1519	0.21	278	512	0.07			
61	1822	0.25	662	1519	0.21	21	478	0.07			
154	1817	0.25	252	1514	0.21	635	460	0.06			

En la **Tabla 8** se muestran las correlaciones entre los 3 parámetros de los ítems y su tasa de exposición. Como era de esperar, los ítems más expuestos son los de mayor parámetro *a* y menor parámetro *c*, dado que son los ítems con mayor nivel de información. Existe también una correlación significativa entre los parámetros de dificultad y las tasas de exposición, que tiene que ver directamente con los niveles de rasgo de la muestra que ha respondido a eCAT: los ítems más presentados tienden a ser difíciles porque el algoritmo adaptativo selecciona ítems de dificultad apropiada al nivel de inglés que tiene la persona evaluada.

Tabla 8
Correlaciones de Pearson entre los parámetros de los ítems y su Tasa de Exposición.

	a	***b***	***c***	***T.E.***
a	1	.170(*)	-.366(**)	.636(**)
b	.170(*)	1	-.138	.356(**)
c	-.366(**)	-.138	1	-.345(**)
T.E.	.636(**)	.356(**)	-.345(**)	1

* La correlación es significante al nivel .05 (bilateral).
** La correlación es significativa al nivel .01 (bilateral).

En la **Tabla 9** se proporciona información comparativa en relación a las propiedades psicométricas esperadas según los estudios de simulación (Abad, Olea, Ponsoda y Ximenez, 2004) y los resultados empíricos. En cuanto a las tasas de exposición, los resultados son bastante similares (se esperan pequeñas diferencias puesto que la distribución del rasgo en los estudios de simulación no coincide con la distribución del rasgo en las aplicaciones empíricas). La sobreexposición de los ítems (ítems con tasas mayores de 0.15) es menor en los datos empíricos probablemente porque el número de ítems disponibles para niveles altos es mayor. Sin embargo, cabe resaltar que numerosos ítems (un 23.9 %) muestran tasas de exposición superiores a 0.25 lo que puede tener que ver con una inadecuada periodicidad en la actualización de los parámetros de exposición. La *tasa de solapamiento*, es decir la proporción de ítems que comparten como media dos examinados, es de 0.222. La mínima tasa de solapamiento posible, si los ítems se aplicaran aleatoriamente, es 0.152.

Tabla 9
Comparativa resultados esperados vs empíricos.
Tasas de exposición.

T.E.	***Tasa de exposición (Abad, Olea, Ponsoda, 2004) Tabla 1. T.Max, 25%***	***Datos empíricos***
0.00-0.05	27.4	26.9
0.05-0.10	9.6	8.6
0.10-0.15	5.6	11.2
0.15-0.20	10.2	6.6
0.20-0.25	47.2	22.8
0.25-0.30	0.0	23.9

En cuanto a las propiedades psicométricas del test (ver **Tabla 10**), el funcionamiento es algo mejor del esperado en los estudios de simulación (puesto que el ajuste entre la distribución del nivel de rasgo y la distribución de la dificultad de los ítems es mayor).

Tabla 10
Porcentaje de personas que cumplen el criterio de parada (error típico menor que 0.30) y coeficiente de fiabilidad para 30 ítems aplicados.

	(Abad, Olea, Ponsoda, 2004). Tabla 2. T.Max, 25%	*Datos empíricos*
%	93%	97%
Coeficiente de fiabilidad	0.94	0.96

8. ESTUDIO SOBRE EL DETERIORO DE LOS PARÁMETROS.

8.1. Comparación de distintos modos de calibración

Se realizaron 3 calibraciones distintas de los 197 ítems: una en el grupo de referencia ***R*** (los 3208 sujetos que respondieron a la versión de papel y lápiz, es decir, las estimaciones de parámetros que actualmente se aplican en eCAT), otra en el grupo focal ***F*** (los 7254 sujetos que respondieron a eCAT) y una tercera incluyendo las respuestas de la muestra total de 10462 sujetos (vamos a llamarla ***R+F***).

En cada caso, para calibrar en la misma métrica los ítems de todos los subtests se utilizó un diseño de calibración concurrente, en el que las respuestas a ítems no aplicados a los sujetos se consideran como *datos perdidos* (Hanson y Béguin, 2002). Para estimar los parámetros de los ítems en las nuevas muestras **(F y R+F**) se utilizó el programa MULTILOG 7.0 (Thissen, Chen y Bock, 2007). En esta ocasión, se utilizaron distribuciones previas similares a las utilizadas con BILOG en el citado estudio previo (Olea, Abad, Ponsoda y Ximenez, 2004). Se utilizo una distribución previa *normal* para el parámetro *a* (media =1.289; desviación típica = .205), normal para el parámetro *b* (media = 0.203; desviación típica = 2) y *logit-normal* para el parámetro *c* (media = .207; desviación típica = .023). Puesto que MULTILOG no permite tratar las omisiones como fraccionalmente correctas (ver Abad, Olea, Ponsoda, Ximénez, y Mazuela, 2004), *se sustituyeron estas por una respuesta aleatoria* (acierto con probabilidad 0.25).

Se comprobó que BILOG y MULTILOG proporcionan idénticos resultados para la muestra de datos original bajo las anteriores especificaciones. Las correlaciones entre las estimaciones obtenidas con BILOG y MULTILOG en la muestra original, ***R***, eran .983, .998 y .950 para los parámetros *a, b* y *c* respectivamente, lo que indica que el cambio en el programa utilizado supone un cambio mínimo en los parámetros obtenidos.

Al aplicar MULTILOG a la muestra *F* debe señalarse que se calibraron los 197 ítems. Puesto que algunos ítems no habían sido aplicados a ninguna persona, la calibración resultaría imposible. Por ello, se añadieron 4 sujetos ficticios con un patrón de respuestas aleatorias:

100010001000….
010001000100….
001000100010….
000100010001….

La repercusión en la estimación debe ser muy pequeña (i.e., la muestra se compone de 7254 sujetos). Al realizar esta operación se obtuvo la convergencia (máximo cambio de parámetros entre ciclos, 0.00089). Los parámetros que proporciona el programa para los ítems que no han sido aplicados en eCAT son los que se establecen en las distribuciones previas bayesianas (a = 1.29, b = 0.00 y c = 0.207). Los errores típicos de estimación para los parámetros en esos ítems son justamente los que indican las distribuciones previas (0.35 para *a*, 2 para *b y* 0.023 para *c*).

Puesto que los parámetros se han estimado en distintas muestras (***R***, ***F*** y ***R+F***) que además 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:*

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

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 estos. Las constantes de equiparación obtenidas con el programa ST (Hanson and Zeng, 1995) aparecen en la **Tabla 11**.

Tabla 11
Constantes de equiparación.

New	*Old*[(1)]	M_1	M_2	*Ítems equiparados*
Muestra Completa ***(R+F)***	*Muestra Lápiz y papel* ***(R)***	.917	.441	197
Muestra eCAT ***(F)***	*Muestra Lápiz y papel* ***(R)***	.834	.637	159[(2)]

[(1)]Los parámetros "*Old"* de los ítems son los parámetros obtenidos previamente con BILOG, es decir, los que están implementados actualmente en eCAT.
[(2)]Para obtener las constantes, se seleccionan los 159 ítems que han sido aplicados al menos a 200 personas.

Puede observarse que en la Tabla 11 M_2 (0.637) y M_1 (0.834) indican la media y la desviación típica del rasgo latente en el grupo *F*.

En la **Tabla 12** se incluyen los datos descriptivos de los 3 parámetros estimados en las 3 diferentes muestras. Interesa fundamentalmente la comparación entre los datos de ***R*** y ***R+F***, dado que finalmente incorporaremos toda la información disponible para actualizar los parámetros del banco. En promedio, se observan escasas discrepancias entre la media de los parámetros estimados en ambas condiciones.

Tabla 12
Distribución de los parámetros después de la equiparación.

		N	*Mínimo*	*Máximo*	*Media*	*Desv. típ.*
Muestra Lapiz y papel	***a-R***	197	.43	2.20	1.30	.32
	b-R	197	-2.71	3.42	.23	1.00
	c-R	197	.11	.29	.21	.03
	a-R	159*	.91	2.20	1.39	.29
	b-R	159*	-2.45	3.42	.35	1.02
	c-R	159*	.11	.29	.20	.03
Muestra eCAT	***a-F***	159*	.31	2.36	1.39	.41
	b-F	159*	-1.65	3.57	.32	.95
	c-F	159*	.12	.25	.20	.01
Muestra Completa	***a-R+F***	197	.34	2.24	1.32	.41
	b-R+F	197	-2.76	3.45	.20	.97
	c-R+F	197	.11	.31	.20	.03

*Se seleccionan los 159 ítems que han sido aplicados al menos a 200 personas en eCAT

Las correlaciones entre los parámetros estimados en las 3 diferentes condiciones muestrales se detallan en la **Tabla 13**. Si nos centramos en las comparaciones entre ***R*** y ***R+F***, puede comprobarse una elevada relación lineal entre las estimaciones de cada

parámetro realizadas en ambas condiciones muestrales. Las variaciones en los parámetros *a* y *b* al considerar la muestra aumentada pueden considerarse pequeñas (o, dicho de otra manera, su valor original puede considerarse precisamente estimado). Cuando comparamos las condiciones ***R*** y ***F,*** los resultados son algo distintos. Las correlaciones entre los parámetros son algo menores, especialmente en los parámetros *a* y *c*. Las correlaciones cuando se seleccionan los 135 ítems aplicados a al menos 500 personas fueron muy similares (.649, .936 y .506), lo que indica que tiende a haber cambios en los parámetros *a* y *c*.

Tabla 13
Correlaciones entre los parámetros en distintas muestras.

		Muestra Lápiz y Papel			***Muestra eCAT***		
		a-R	***b-R***	***c-R***	***a-F***	***b-F***	***c-F***
Muestra eCAT	***a-F***	.645*					
	b-F		.940*				
	c-F			.510*			
Muestra Completa	***a-R+F***	.864			.922		
	b-R+F		.974			.989	
	c-R+F			.912			.598

*Las correlaciones están calculadas con los 159 ítems aplicados a al menos 200 personas.

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 ***R+F***; también se obtuvo el valor absoluto de dichas diferencias. Las correlaciones entre estas variables y las tasas de exposición fueron las que aparecen en la **Tabla 14**.

Tabla 14
Correlaciones entre el cambio en los parámetros *a*, *b* y *c* para los 197 ítems (cambio relativo y absoluto, muestras R y R+F) y la Tasa de Exposición (T.E.).

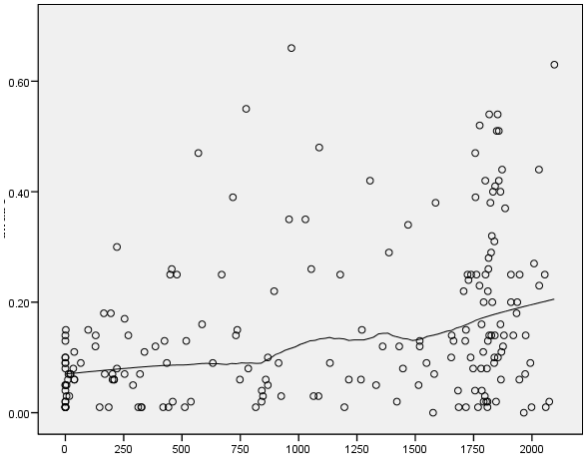
			Cambio relativo			***Cambio absoluto***		
		T.E.	***a***	***b***	***c***	***a***	***b***	***c***
Cambio relativo	***a***	.056	1	-.169(*)	.030			
	b	.030	-.169(*)	1	.368(**)			
	c	.007	.030	.368(**)	1			
Cambio absoluto	***a***	.343(**)				1	.402(**)	.084
	b	.253(**)				.402(**)	1	.235(**)
	c	.171(*)				.084	.235(**)	1

** La correlación es significativa al nivel 0,01 (bilateral).
* La correlación es significante al nivel 0,05 (bilateral).

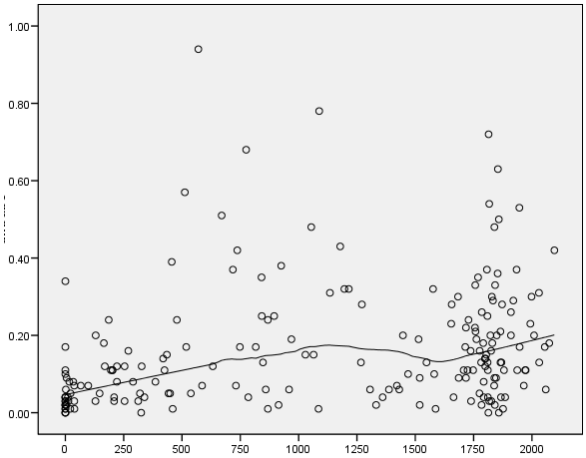
En general, hay relación entre las diferencias absolutas en los parámetros estimados y las tasas de exposición. Parece que las variaciones en los parámetros *a, b* y *c* son mayores cuando disponemos de tamaños muestrales elevados en eCAT para estimarla, aunque estas variaciones no son sistemáticas (unas veces son mayores y otras menores de las estimadas en el grupo ***R***). Esto es lógico ya que a mayor número de aplicaciones en eCAT, mayor importancia cobra la muestra de eCAT en la determinación de los parámetros. El patrón de correlaciones se mantiene cuando se excluyen los ítems de anclaje (que fueron originalmente aplicados a un número mayor de sujetos). En las **figuras 4, 5 y 6** se muestran los cambios absolutos en *a, b* y *c* como función de la tasa de exposición.

Figuras 4, 5 y 6
Cambio absoluto en *a*, *b* y *c* en la muestra R+F [abs(a_{R+F}-a_R)] como función del número de aplicaciones de eCAT.

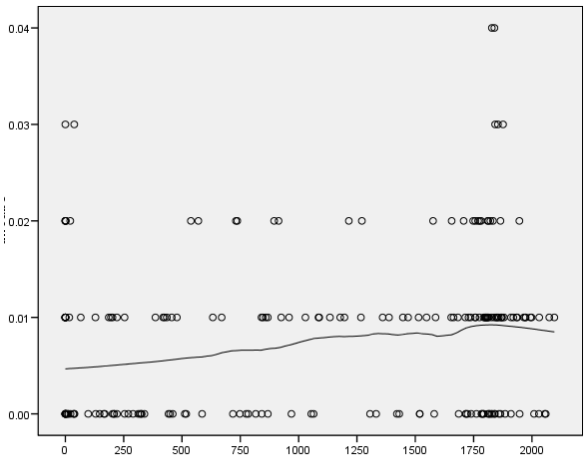
Cambio en a



Cambio en b



Cambio en c



El análisis anterior indica como cambiarían los parámetros al considerar la muestra completa. Los resultados obtenidos al comparar los parámetros de *todos* los ítems en las condiciones ***R*** y ***F*** nos informan más directamente de cómo cambian los parámetros de eCAT en función del número de aplicaciones. Las correlaciones entre las variables de cambio y las tasas de exposición aparecen en la **Tabla 15**.

Tabla 15
Correlaciones entre el cambio en los parámetros *a*, *b* y *c* para los 197 ítems (cambio relativo y absoluto, muestras R y F) y la Tasa de Exposición (T.E.).

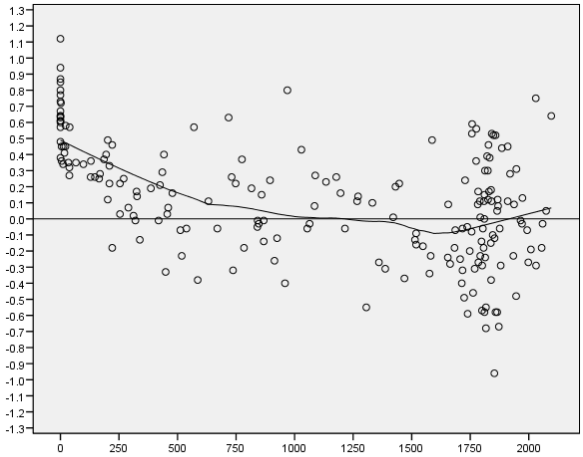
			Cambio relativo			*Cambio absoluto*		
		T.E.	***a***	***b***	***c***	***a***	***b***	***c***
Cambio relativo	***a***	-.504(**)	1	.405(**)	-.330(**)			
	b	-.405(**)	.405(**)	1	-.034			
	c	.246(**)	-.330(**)	-.034	1			
Cambio absoluto	***a***	-.251(**)				1	.501(**)	.083
	b	-.402(**)				.501(**)	1	-.025
	c	.182(*)				.083	-.025	1

** La correlación es significativa al nivel 0,01 (bilateral).
* La correlación es significante al nivel 0,05 (bilateral).

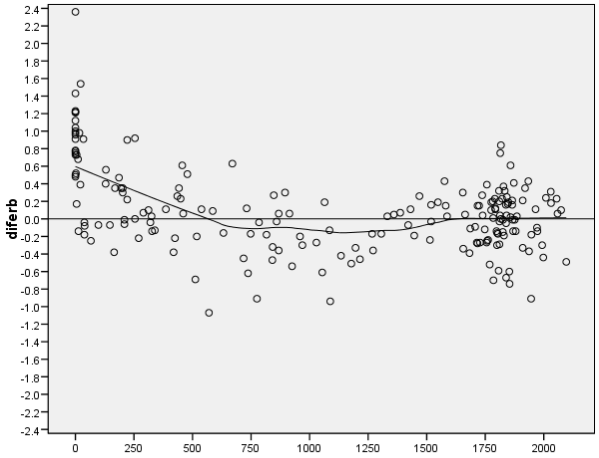
En este caso, se observa relación entre las diferencias (absolutas y relativas) en los parámetros estimados y las tasas de exposición. A mayor tasa de exposición, menores parámetros *a* y *b* y mayores parámetros *c.* En general, a medida que aumenta el número de aplicaciones la diferencia media entre los parámetros en ambos formatos (lápiz y papel y eCAT) se aproxima a 0 (ver **figuras 7, 8 y 9**) y, tal como muestran las correlaciones, el cambio absoluto tiende a ser menor (en *a* y en *b*). Esto es lógico ya que cuando no hay suficientes sujetos para estimar los parámetros los valores de estos convergen a los especificados en las distribuciones previas. Así, en el caso del parámetro *a* se observa inicialmente un sesgo positivo cuando hay pocas aplicaciones (los ítems menos expuestos son justamente los de menor parámetro *a*); a medida, que un ítem es más aplicado el sesgo medio tiende a 0. Para *b* ocurre un patrón similar (los ítems más fáciles son los menos aplicados y en ausencia de aplicaciones, se sobreestima su *b*). En el parámetro *c,* el patrón es algo distinto. A medida que aumenta el número de aplicaciones, los valores de *c* en la muestra de eCAT tienden a haber más cambio en términos absolutos.

Figuras 7, 8 y 9
Cambio *relativo* en *a*, *b* y *c* en la muestra *F* [a_F-a_R] como función del número de aplicaciones de eCAT.

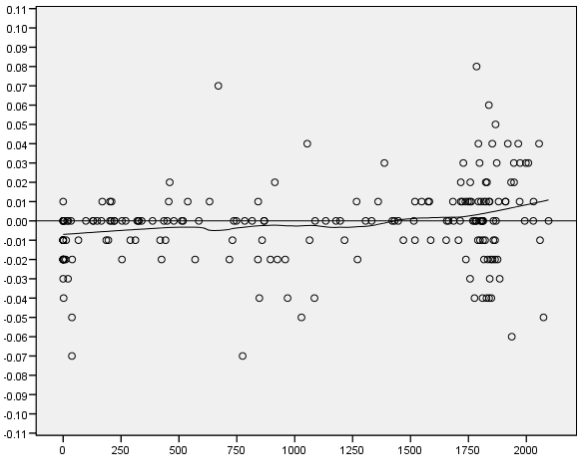
Cambio en a



Cambio en b



Cambio en c



Puede concluirse que cuando el número de aplicaciones de eCAT es alto:

1. No hay cambios *sistemáticos* en los parámetros. A partir de 500 aplicaciones de eCAT, la diferencia media entre los parámetros *a y b* en los dos formatos se aproxima a 0.
2. Hay *cambios* no sistemáticos en los parámetros (especialmente en *a* y *c*), lo que implica la necesidad de que los parámetros de los ítems sean actualizados a los valores obtenidos en la muestra ampliada.
3. Se podría pensar que el hecho de que la media y desviación típica de θ sea 0.637 y 0.834 podrían indicar que los parámetros se han vuelto *más* fáciles y *más* discriminativos y que esas diferencias se han traslado a la distribución del rasgo. Sin embargo, la razón principal por la que un ítem se puede hacer más fácil (la difusión del banco) no parece apoyada por los datos (i.e., no parece que a mayor tasa de exposición los ítems se hagan más fáciles).

8.2. Análisis de la calidad de la calibración con la muestra de eCAT

Una forma sencilla de comprobar en qué medida las correlaciones entre los parámetros son bajas por las condiciones de calibración (p.e., dificultad de las condiciones de estimación por el número de datos perdidos) es estudiar las correlaciones entre los parámetros obtenidos en dos submuestras de los datos de eCAT. Así, se procedió a dividir la muestra en dos mitades equivalentes (muestra 1: número de caso impar; muestra 2: número de caso par). Los datos se muestran en la **Tabla 16.**

Tabla 16
Correlaciones entre parámetros *a, b y c* en dos muestras distintas de eCAT*

	a_2	b_2	c_2
a_1	**.900**	.098	-.599
b_1	.106	**.994**	-.263
c_1	-.534	-.244	**.837**

*para los 143 ítems que han sido aplicados al menos 400 veces en la muestra total de eCAT: es decir, que probablemente han sido aplicados alrededor de 200 veces en cada submuestra.

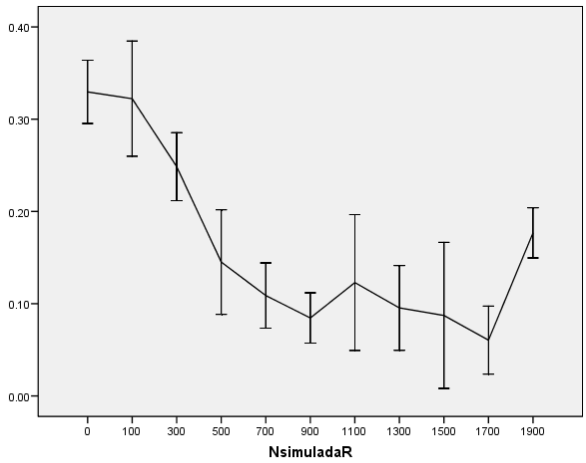
Puede observarse que las correlaciones entre parámetros en las dos muestras son bastante altas. Llama la atención la alta correlación entre los parámetros *a* y *c,* que

también se encuentra entre los parámetros *a* y *c* originales (r_{ac}=-.366). Los ítems con menor parámetro *a* tienden a tener un mayor parámetro *c*.

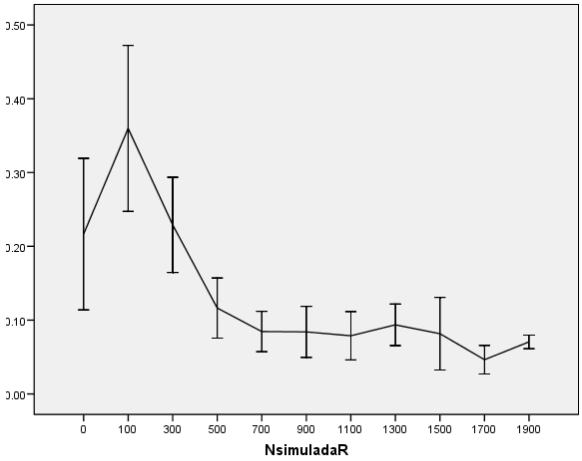
También se realizó un pequeño estudio de simulación. Se simularon respuestas a eCAT con los parámetros originales para 7254 sujetos con una distribución N(0.5, 1). Las **figuras 10, 11 y 12** muestran el error absoluto (diferencia entre parámetro real y estimado) medio como función del número de aplicaciones para *a*, *b* y *c* (en el eje de abcisas se representa el punto medio del intervalo de aplicaciones; por ejemplo, 100 índica ítems aplicados entre 1 y 200 veces).

Figuras 10, 11 y 12
Diferencia absoluta media (e intervalo de confianza) en *a*, *b* y *c* en la muestra *F* [|a_estimada-a_areal|] como función del número de aplicaciones de eCAT (negro: simulación).

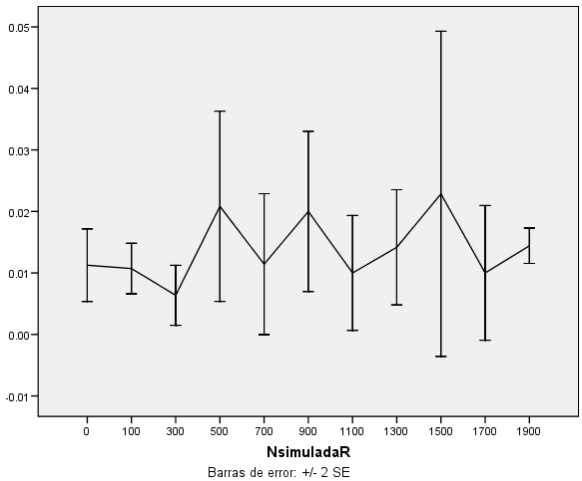
Diferencia en a



Diferencia en b



Diferencia en c



Lo anterior indica que la estimación del parámetro *c* es muy pobre. Parece que para los ítems aplicados más de 400 veces los valores medios del cambio absoluto son muy parecidos. Puede observarse que en los ítems más aplicados hay una peor estimación del parámetro *a;* esto se debe a que esos ítems son los que poseen mayor parámetro de discriminación, que es más difícil de estimar. Las correlaciones entre parámetros reales y estimados para ítems aplicados más de 400 veces se encuentran en la **Tabla 17**.

Tabla 17
Correlaciones entre parámetros *a, b y c reales y estimados**

	a_2	b_2	c_2
a	**.827**	.045	-.433
b	-.118	**.999**	-.218
c	-.415	-.131	**.567**

*para los 151 ítems que han sido aplicados al menos 400 veces en la simulación

Obsérvese que los valores en la diagonal son, similares, en el caso de los parámetros *a* y *b*, a los valores diagonales de la Tabla 16 al cuadrado (.81 y .99). Esto puede indicar que la alta correlación en el parámetro *c* al utilizar dos sub-muestras es un artefacto metodológico de algún tipo y no refleja la calidad de la estimación del parámetro *c.*

Los valores de RMSE y el sesgo en función del número de aplicaciones del ítem aparecen en la **Tabla 18.** Puede observarse que el parámetro *b* se estima bastante bien incluso en muestras pequeñas. La estimación mejora al aumentar la muestra. Con respecto al parámetro *b,* se produce un sesgo negativo y una incorrecta estimación de los ítems más aplicados que son los que tienen parámetros *a* más altos. Finalmente, el parámetro *c* se estima bastante mal pues el RMSE(c) no es menor que el error típico si no hubiera aplicaciones, salvo cuando el número de aplicaciones es muy alto (mayor de 1500).

Tabla 18
RMSE para los parámetros *a, b y c* según número de aplicaciones.

Nº aplicaciones	***N***	***RMSE(a)***	***RMSE(b)***	***RMSE(c)***	***Sesgo(a)***	***Sesgo(b)***	***Sesgo(c)***
400-1000	39	.141	.116	.028	.062	-.071	-.011
1000-1500	28	.134	.098	.026	-.062	-.063	-.004
1500-2000	84	.208	.081	.019	-.152	-.059	.005
Desviación típica de la distribución previa		.35	2.00	.023			

8.3. Estudio de evaluación del DIF

Hemos descrito las razones que pueden llevar a un cambio de parámetros de los ítems tras las aplicaciones adaptativas. En el apartado anterior se observaron cambios en los parámetros *a* y *c*. Estos cambios no son los esperables dada la variación muestral, especialmente para *a*. Por lo tanto, pasamos a ver en qué ítems se producen las variaciones y si merecen consideración algunas de las razones hipotéticas descritas (p.e., difusión de los ítems del banco).

El estudio de estas posibles variaciones se realizó mediante una estrategia de evaluación del Funcionamiento Diferencial de los Ítems (DIF). En concreto, se aplicó un procedimiento paramétrico basado en la TRI, usando el marco DFIT (*Differential Functioning of Ítems and Test*), que permite evaluar el funcionamiento diferencial del ítem y el funcionamiento diferencial del test, simultáneamente. En nuestro caso, sólo estamos interesados en el funcionamiento diferencial de los ítems puesto que el test que se aplica a cada persona es distinto y el número de aciertos en el test no reflejará con precisión su nivel de rasgo. Estudiar cómo se traducen exactamente los cambios de los parámetros de los ítems en sesgos en la estimación de θ es abordado en un estudio de simulación posterior.

Una vez equiparados los parámetros de TRI de ambos grupos ***R*** y ***F*** se puede obtener el siguiente indicador para cada ítem:

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

donde NCDIF indica el promedio de las diferencias cuadraticas entre las CCIs de ambos grupos a través del rasgo (θ). Las integrales se aproximan por puntos de cuadratura y se asume que el rasgo se distribuye normalmente. El NCDIF es una medida muy similar a otras medidas 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:

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

Se obtuvieron los valores del NCDIF para los ítems aplicados a más de 200 evaluados. Los resultados se muestran en la **Tabla 19**, donde se observa 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 (o ventaja promedio) no fueron significativas (r = 0.107 para NCDIF y r = 0.147 para la ventaja promedio; p > .05), lo cual indica que los ítems con más variación en sus parámetros no son los que más se presentan en eCAT. Este resultado es muy relevante, porque indica, de nuevo, que probablemente no ha habido problemas de transmisión de ítems entre los evaluados por eCAT. La media de la V.P. es próxima a cero (-0.003). Para el resto de los ítems contrastados (101 ítems) se puede considerar que sus parámetros son estables (NCDIF < .006)[1].

[1] 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.

Tabla 19
ítems con NCDIF mayor o igual a 0.006.

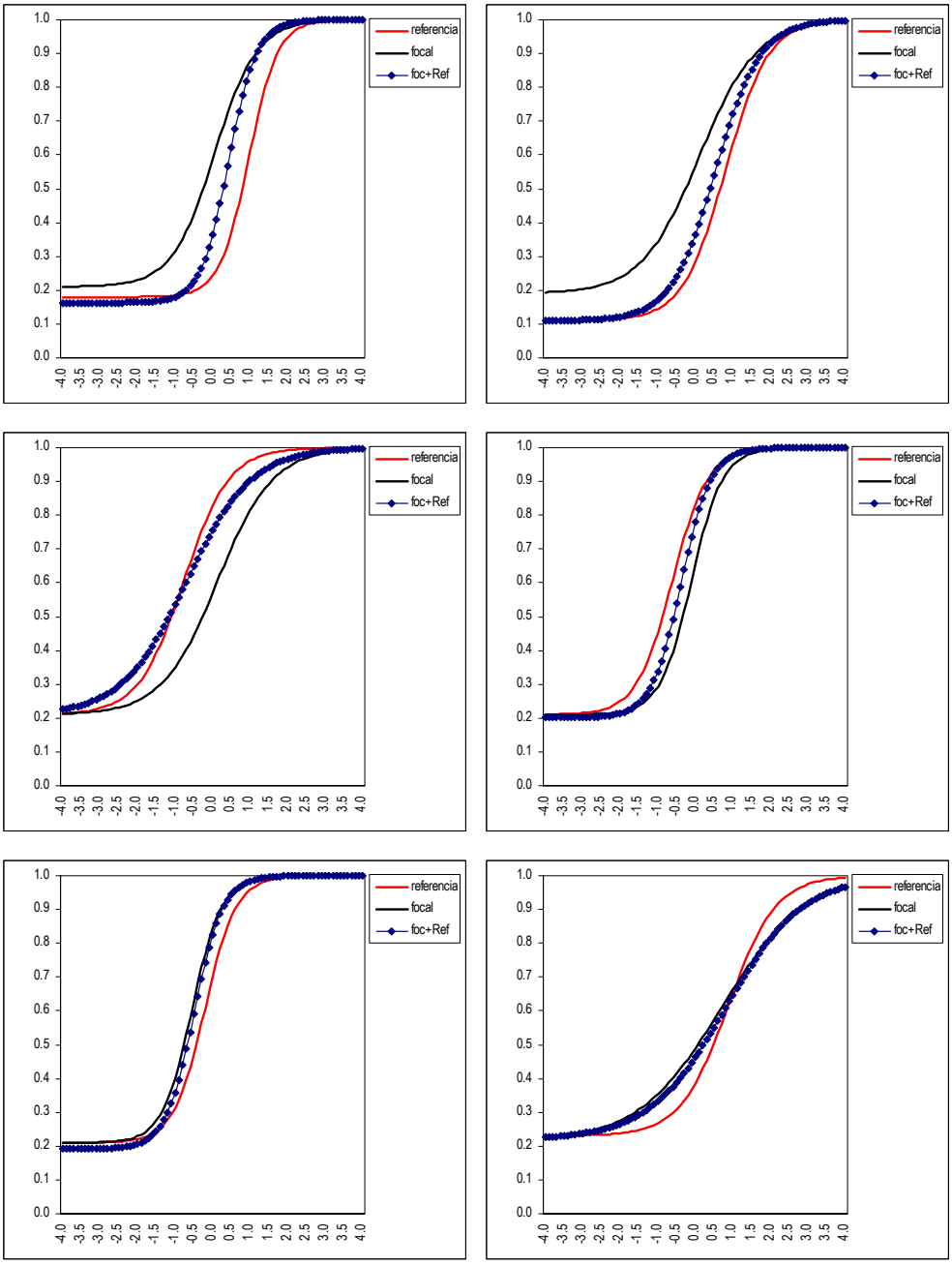
		Lapiz y papel (R)			eCAT (F)				
iditem	NºAp	a	b	c	a	b	c	V.P.*	NCDIF
79	1946	1.53	0.96	0.18	1.05	0.05	0.21	0.239	0.071
607**	1785	1.04	0.80	0.11	0.77	0.10	0.19	0.213	0.050
494	1853	1.75	0.92	0.17	0.79	0.32	0.21	0.195	0.049
95	221	0.98	-0.77	0.21	0.80	0.13	0.21	-0.187	0.040
148**	1757	1.62	0.02	0.19	2.15	0.41	0.16	-0.130	0.026
369	1828	1.90	0.77	0.16	2.36	1.14	0.12	-0.119	0.022
479	1854	1.03	1.26	0.22	0.91	0.52	0.20	0.126	0.022
232	456	1.36	-1.22	0.20	1.39	-0.61	0.21	-0.114	0.022
121	1769	1.50	-0.07	0.21	1.19	-0.59	0.21	0.119	0.021
164	1088	1.38	2.24	0.20	1.65	1.30	0.20	0.078	0.021
60	1998	1.40	0.78	0.18	1.13	0.34	0.21	0.129	0.021
21	478	1.25	-0.62	0.21	1.41	-0.11	0.21	-0.118	0.020
344	1807	1.04	-0.11	0.21	0.86	-0.70	0.21	0.114	0.017
350	1839	1.05	1.17	0.26	0.67	0.50	0.22	0.115	0.016
374	1858	1.73	0.83	0.20	1.15	1.44	0.20	-0.077	0.016
384	1215	1.37	-0.02	0.22	1.31	-0.48	0.21	0.107	0.016
154	1817	0.99	0.74	0.20	0.31	1.58	0.21	0.016	0.016
209	1576	1.47	-0.39	0.21	1.13	0.04	0.22	-0.104	0.015
58	1833	1.58	0.22	0.21	1.96	0.54	0.19	-0.101	0.015
621**	2096	1.27	1.24	0.20	1.91	0.75	0.20	0.070	0.014
358	775	1.29	2.31	0.28	1.66	1.40	0.21	-0.004	0.013
598	1683	1.49	-0.37	0.20	1.31	-0.76	0.21	0.087	0.012
30**	442	0.91	0.71	0.22	1.31	1.06	0.21	-0.097	0.012
427	1800	1.06	0.19	0.22	0.49	-0.11	0.21	0.054	0.012
415	254	0.94	-2.45	0.21	0.97	-1.53	0.21	-0.075	0.012
518	737	1.07	2.26	0.21	0.75	1.64	0.21	0.087	0.011
189	202	1.08	0.65	0.20	1.57	1.00	0.21	-0.084	0.011
536**	895	1.34	0.22	0.23	1.58	0.52	0.21	-0.090	0.011
24	1814	1.06	1.02	0.25	0.82	1.77	0.25	-0.073	0.011
617	1656	1.41	-0.21	0.21	1.50	-0.55	0.21	0.085	0.010
50	1654	1.70	-0.56	0.20	1.46	-0.26	0.19	-0.084	0.010
535	1776	1.71	0.11	0.22	2.27	0.30	0.18	-0.079	0.010
10	1872	1.53	0.51	0.19	0.86	0.92	0.22	-0.028	0.010
240	1134	1.00	-0.97	0.21	1.23	-1.39	0.21	0.082	0.010
157	1054	1.62	2.44	0.18	1.56	1.83	0.22	0.073	0.009
67	1849	1.63	0.71	0.22	2.15	0.90	0.18	-0.082	0.009
296	2030	1.32	1.22	0.18	2.07	1.53	0.18	-0.067	0.009
155	847	1.37	0.19	0.25	1.34	0.46	0.21	-0.087	0.009
86	1791	2.13	0.14	0.19	2.24	0.37	0.18	-0.072	0.009
116	1196	1.37	-0.64	0.20	1.53	-0.97	0.20	0.075	0.009
12	1469	1.06	-1.13	0.21	0.69	-0.87	0.20	-0.083	0.009
69**	1842	1.22	1.28	0.24	1.75	1.53	0.21	-0.080	0.009
519	1306	1.01	-0.76	0.21	0.46	-0.93	0.21	-0.039	0.009
11	1717	1.79	-0.22	0.21	1.47	-0.50	0.22	0.069	0.008
169	1934	1.49	0.79	0.20	1.26	1.22	0.22	-0.052	0.008
61**	1822	1.60	0.11	0.21	1.99	0.34	0.20	-0.071	0.008
107	1728	1.28	0.63	0.18	1.52	0.36	0.21	0.077	0.008
412	842	1.30	-0.43	0.20	1.29	-0.75	0.21	0.074	0.008
438	419	1.02	-0.11	0.22	1.01	-0.49	0.21	0.078	0.007
1**	1811	2.19	0.49	0.26	1.61	0.20	0.22	0.054	0.007
462	1762	1.55	0.00	0.20	1.09	-0.24	0.21	0.060	0.007
104	1920	1.50	0.86	0.17	1.78	1.21	0.21	-0.041	0.007
422	449	0.97	-0.98	0.21	0.64	-0.75	0.21	-0.073	0.007
80	1809	1.03	0.27	0.21	1.14	0.59	0.21	-0.074	0.007
85	1993	1.45	0.92	0.20	1.38	0.62	0.20	0.066	0.007
197	435	1.28	-0.11	0.21	1.57	0.15	0.21	-0.067	0.007
285	1910	1.25	1.07	0.20	1.36	0.74	0.21	0.066	0.007
375	959	0.93	0.87	0.23	0.53	0.67	0.21	0.062	0.006
48	221	1.16	0.48	0.21	1.62	0.70	0.21	-0.064	0.006
Media	1405							0.003	

*Ventaja promedio: positiva: a favor del grupo focal; negativa: a favor del grupo de referencia.
**ítem de anclaje

En las siguientes gráficas se muestran los resultados para ítems con distinto grado de FDI (0.071, 0.050, 0.040, 0.020, 0.010 y 0.006).

Figuras 13, 14, 15, 16, 17 y 18

6 ítems con FDI (79, 607, 95, 21, 617 y 375)



Finalmente, se recalcularon las constantes de equiparación M_1 y M_2 después de eliminar los ítems con NCDIF mayor o igual que 0.020 (ya que la inclusión de ítems con FDI podría haber afectado a la equiparación). Las nuevas constantes eran muy similares a las

iniciales ($M_1 = 0.832$; $M_2 = 0.637$) por lo que se tomaron los resultados descritos como definitivos.

8.4. Repercusión en θ estimada del cambio de parámetros

En TAIs es 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 distinto test por lo tanto la puntuación esperada en el test tiene poco valor práctico. Por ello, en el presente informe se estudia, mediante simulación, la diferencia esperada en la θ *estimada* como función de θ.

Estudio de simulación: Se simularon las respuestas de 7254 sujetos de una distribución normal N(0.5, 1). A cada evaluado se le aplica eCAT considerando los parámetros *originales* en la selección de ítems y en la estimación de θ *a través del TAI.* Es decir, se simula el funcionamiento del TAI operativo tal como está implementado. Las respuestas de los sujetos a eCAT se simulaban según los *nuevos* o "verdaderos" parámetros de los ítems[2]. Posteriormente, se estudia el valor esperado de la θ *estimada* según los parámetros utilizados en la estimación de θ al final (con los 30 ítems ya aplicados):

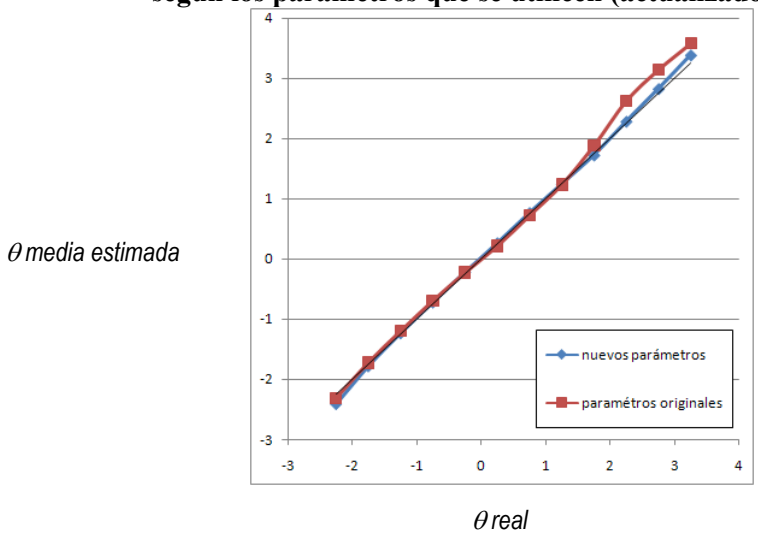
- *Parámetros de eCAT (los utilizados en el TAI).*
- *Nuevos parámetros (los utilizados en la generación de las respuestas).*

Resultados: El sesgo al estimar θ producido por utilizar los parámetros originales (en vez de los parámetros actualizados) es muy pequeño. En la **Figura 19** se muestra el valor esperado de la θ *estimada:*

[2] Los nuevos parámetros eran estimados exclusivamente con la muestra de eCAT para aquellos ítems para los que se disponía de más de 400 respuestas (143 ítems). En el resto de los ítems (54 ítems) se mantienen siempre los parámetros originales.

La correlación entre ambos conjuntos de θ con la θ real es muy alta (r = 0.973 y 0.968, según se utilicen los nuevos parámetros o los originales); por lo tanto, puede decirse que la repercusión de la actualización de parámetros es muy pequeña. Puede observarse que con los parámetros originales tiende a producirse un valor esperado más alto en los niveles *altos*.

Figura 19
Gráfico de dispersión que muestra la relación entre las zetas obtenidas según los parámetros que se utilicen (actualizados o no).



También se calculó el gráfico de dispersión entre las θ estimadas con ambos conjuntos de parámetros (ver **figura 20**). La correlación fue también muy alta en este caso (r = 0.993). Es también claro 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) y que la clasificación de θ puede modificarse (ver **Tablas 20** y **21**), lo que indica la conveniencia de cambiar los parámetros.

Figura 20
Gráfico de dispersión que muestra la relación entre las zetas obtenidas según los parámetros que se utilicen (calibrados solo con eCAT u originales).

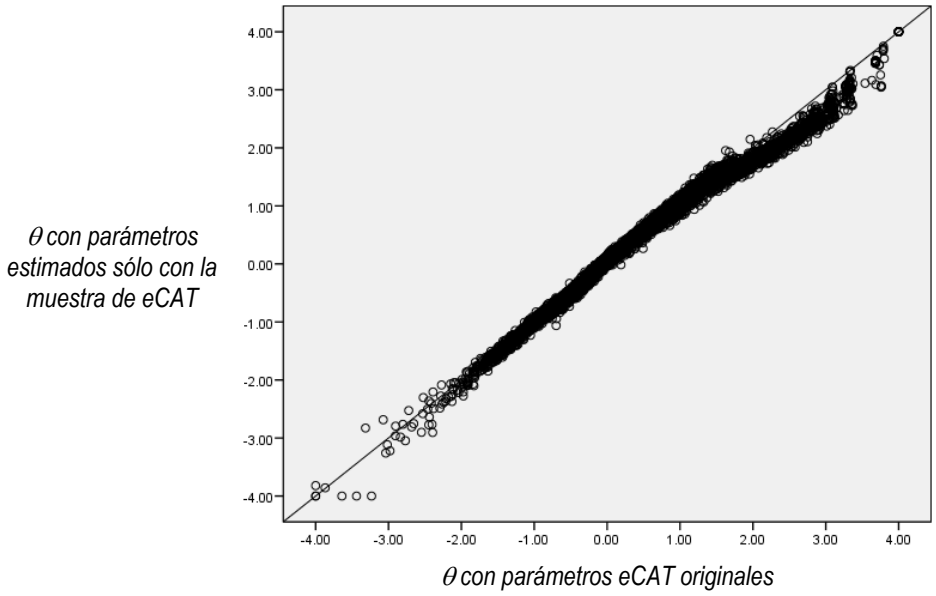


Tabla 20
"Sobreestimación de θ" por usar los parámetros desactualizados

Decil de θ con los parámetros originales	N	Mínimo	Máximo	Media	Desv. típ.
1	725	-.49	.77	.02	.09
2	724	-.18	.36	.03	.06
3	727	-.16	.13	-.03	.05
4	724	-.19	.21	-.06	.04
5	728	-.22	.13	-.07	.05
6	724	-.22	.23	-.06	.06
7	724	-.21	.20	-.04	.07
8	727	-.29	.26	-.04	.08
9	725	-.33	.41	.08	.12
10	726	-.01	.71	.32	.13

Tabla 21
Diferencias en la clasificación de θ por deciles

		θ con parámetros estimados sólo con la muestra de eCAT									
		1	2	3	4	5	6	7	8	9	10
θ con parámetros eCAT originales	1	693	32	0	0	0	0	0	0	0	0
	2	32	653	39	0	0	0	0	0	0	0
	3	0	41	639	47	0	0	0	0	0	0
	4	0	0	48	630	46	0	0	0	0	0
	5	0	0	0	47	615	66	0	0	0	0
	6	0	0	0	1	62	595	66	0	0	0
	7	0	0	0	0	0	67	582	75	0	0
	8	0	0	0	0	0	0	77	585	65	0
	9	0	0	0	0	0	0	0	65	622	38
	10	0	0	0	0	0	0	0	0	38	688

9. PROPUESTA EN RELACIÓN AL MANTENIMIENTO (O NO) DE LOS PARÁMETROS DE LOS ÍTEMS OPERATIVOS

A partir de todo lo anterior (cambios claros en los parámetros *a* y *c*, presencia de ítems con FDI y repercusión leve de los cambios en los niveles altos), nuestra propuesta es que se modifiquen los parámetros de los 197 ítems, incluyendo en eCAT los nuevos parámetros estimados con la muestra total de 10462 sujetos. Esto permite actualizar los parámetros de los ítems de forma que el peso de la muestra de eCAT en el valor del nuevo parámetro dependa del número de aplicaciones, lo cual parece conveniente. De esta manera, si un ítem se ha aplicado poco en eCAT mantendrá su valor original.

¿Deberían eliminarse del banco los ítems no aplicados? Es obvio que los ítems no administrados no resultan muy informativos para los niveles de inglés de los sujetos que responden a eCAT, pero proponemos su mantenimiento en el banco por si en un futuro se aplica la prueba a personas de nivel medio-bajo. Por otra parte, de los estudios realizados se concluye que no existe un problema serio de transmisión de ítems, con lo cual hemos decidido mantener los ítems con mayores tasas de exposición. Más bien parece que los cambios en los parámetros pueden deberse a cambios inespecíficos en las condiciones de aplicación. Por tanto, parece razonable actualizar los parámetros sin eliminar los ítems. Los parámetros a incorporar se muestran en la **Tabla 21**. Una alternativa, en relación *a los ítems de anclaje con DIF*, es tomar, para esos ítems, los parámetros estimados con la muestra de eCAT ya que al haber sido aplicados originalmente a una muestra muy grande sus parámetros se *actualizan* muy lentamente al incorporar la muestra de eCAT (ver **Tabla 22**). En realidad, la diferencia entre una tabla y otra es muy pequeña por lo que nuestra recomendación sería usar la **Tabla 21**.

Tabla 21

Alternativa A: Nuevos parámetros para los ítems de eCAT (D = 1.7).

Id	a	b	c
1	1.97	0.38	0.26
2	0.64	-2.16	0.21
3	1.75	0.45	0.20
8	1.38	-0.30	0.20
9	1.29	-0.88	0.21
10	1.09	0.79	0.20
11	1.64	-0.44	0.21
12	0.72	-1.03	0.20
15	1.22	-0.28	0.21
18	1.82	0.25	0.17
20	2.01	-0.09	0.16
21	1.50	-0.38	0.20
22	1.30	-0.03	0.20
23	0.80	0.46	0.19
24	0.78	1.74	0.26
29	0.91	0.13	0.21
30	0.90	0.76	0.22
33	1.94	0.82	0.14
35	0.95	-0.10	0.22
41	2.04	0.48	0.19
42	1.02	-0.85	0.21
48	1.24	0.60	0.22
50	1.60	-0.33	0.19
53	1.39	-0.37	0.22
57	1.36	0.22	0.21
58	1.98	0.51	0.23
60	1.41	0.48	0.17
61	1.98	0.31	0.23
62	1.16	1.20	0.18
63	1.12	0.37	0.23
66	1.75	0.90	0.23
67	2.14	0.91	0.23
69	1.63	1.61	0.27
73	0.99	0.19	0.23
74	1.98	2.09	0.21
76	1.22	1.13	0.30
79	1.59	0.43	0.16
80	1.04	0.52	0.22
84	1.83	0.36	0.16
85	1.54	0.69	0.19
86	2.24	0.32	0.19
88	1.20	-1.22	0.21
89	0.93	-0.13	0.23
95	0.68	-0.69	0.21
99	0.76	-2.76	0.21
100	0.47	-0.32	0.21
102	1.08	-0.06	0.19
104	1.64	1.15	0.18
105	1.55	-0.85	0.20
107	1.52	0.39	0.17
109	1.81	0.21	0.19
111	1.24	-0.40	0.23
112	1.27	0.37	0.26
115	0.57	-0.01	0.21
116	1.38	-0.96	0.19
120	2.08	0.27	0.16
121	1.51	-0.42	0.19
123	1.27	-0.45	0.23
124	0.89	-0.98	0.21
125	1.26	-0.46	0.21
132	1.48	0.13	0.19
135	1.39	0.02	0.19
145	0.97	-0.32	0.20
148	2.09	0.35	0.17
153	1.42	0.87	0.18
154	0.45	1.28	0.20
155	1.40	0.32	0.26
157	1.88	1.96	0.18
161	1.04	1.36	0.16
164	1.86	1.46	0.19
167	2.01	1.93	0.14
169	1.31	1.16	0.21
171	0.79	-1.01	0.21
176	1.13	1.25	0.19
179	1.18	0.64	0.22
181	1.22	-0.22	0.21
184	2.04	0.44	0.18
188	1.31	-0.33	0.21
189	1.01	0.76	0.21
190	0.90	1.45	0.19
192	0.58	-0.50	0.21
196	1.35	-0.05	0.22
197	1.37	0.04	0.22
198	1.65	1.20	0.16
201	1.39	1.51	0.20
207	0.77	-0.35	0.22
209	1.47	-0.07	0.23
220	1.55	1.08	0.21
224	2.14	0.20	0.20
226	2.13	0.57	0.16
231	1.42	-0.10	0.21
232	1.62	-0.83	0.21
233	0.90	-0.24	0.21
237	1.26	-0.53	0.21
238	1.03	-1.18	0.21
240	1.09	-1.28	0.20
243	1.35	0.69	0.22
247	1.24	0.31	0.21
248	1.34	0.63	0.21
251	1.21	0.63	0.23
252	1.37	-0.69	0.20
253	1.43	1.63	0.26
257	1.95	0.73	0.23
258	0.79	-0.63	0.21
274	1.46	1.87	0.22
276	0.81	0.09	0.20
277	1.26	1.96	0.22
278	1.21	2.34	0.22
280	1.40	1.76	0.24
281	0.88	1.04	0.19
285	1.45	0.81	0.20
296	1.76	1.53	0.18
297	1.72	-0.05	0.18
300	1.02	0.74	0.20
307	1.95	0.00	0.23
308	1.23	1.99	0.24
313	1.29	-1.19	0.21
326	0.96	-0.02	0.22
330	1.47	0.43	0.23
339	0.99	-0.77	0.21
343	1.01	-1.45	0.21
344	0.96	-0.48	0.20
348	1.44	-0.54	0.19
350	0.74	0.69	0.22
358	1.84	1.63	0.28
359	1.26	0.27	0.21
363	1.20	0.96	0.21
364	0.77	0.72	0.20
367	0.67	0.35	0.23
369	2.22	1.07	0.12
372	1.78	-0.66	0.20
374	1.22	1.33	0.20
375	0.58	0.81	0.22
377	1.48	0.97	0.19
381	1.86	0.39	0.17
384	1.43	-0.34	0.20
390	1.54	-0.59	0.20
391	1.19	-1.66	0.21
393	2.17	0.43	0.16
394	1.26	0.48	0.19
403	1.47	0.09	0.21
406	1.62	0.41	0.20
412	1.26	-0.68	0.20
415	0.77	-2.33	0.21
416	1.15	0.41	0.22
421	1.54	-1.51	0.20
422	0.72	-0.93	0.21
424	0.96	-1.21	0.21
427	0.64	0.07	0.21
430	1.32	-0.58	0.20
433	1.25	0.64	0.22
438	1.01	-0.25	0.21
442	1.39	1.13	0.18
447	1.36	0.32	0.31
454	1.38	-0.24	0.22
462	1.30	-0.19	0.20
479	0.93	0.63	0.19
480	1.16	-0.31	0.22
494	1.21	0.56	0.18
498	1.80	1.40	0.13
505	0.73	-0.50	0.21
518	0.92	1.84	0.19
519	0.59	-0.82	0.21
524	1.77	1.89	0.14
535	2.23	0.27	0.20
536	1.56	0.47	0.25
545	1.69	2.48	0.17
550	1.63	0.32	0.21
554	1.55	1.88	0.19
567	0.95	-0.09	0.23
571	0.95	2.10	0.16
584	1.08	1.65	0.22
590	1.23	-0.61	0.23
591	1.33	-0.85	0.21
593	0.83	0.50	0.22
595	1.43	-1.06	0.20
596	1.24	-0.43	0.21
597	1.07	-0.56	0.20
598	1.45	-0.67	0.19
599	1.45	-0.76	0.21
606	0.34	-2.06	0.21
607	1.00	0.54	0.11
612	0.65	-0.24	0.21
615	1.50	0.16	0.20
617	1.55	-0.49	0.19
621	1.90	0.82	0.21
623	1.13	-0.97	0.22
625	1.46	-1.32	0.20
628	1.05	3.45	0.12
635	1.24	0.13	0.19
636	2.06	-0.74	0.19
641	1.25	-1.30	0.22
644	0.93	-0.57	0.21
653	1.73	0.54	0.24
661	0.87	-0.55	0.22
662	0.90	-0.52	0.20
663	1.49	0.07	0.21

Tabla 22
Alternativa B: Nuevos parámetros para los ítems de eCAT (D = 1.7)

Id	a	b	c
1	1.61	.20	.22
2	0.64	-2.16	0.21
3	1.75	0.45	0.20
8	1.38	-0.30	0.20
9	1.29	-0.88	0.21
10	1.09	0.79	0.20
11	1.64	-0.44	0.21
12	0.72	-1.03	0.20
15	1.22	-0.28	0.21
18	1.82	0.25	0.17
20	2.01	-0.09	0.16
21	1.50	-0.38	0.20
22	1.30	-0.03	0.20
23	0.80	0.46	0.19
24	0.78	1.74	0.26
29	0.91	0.13	0.21
30	1.31	1.06	.21
33	1.94	0.82	0.14
35	0.95	-0.10	0.22
41	2.04	0.48	0.19
42	1.02	-0.85	0.21
48	1.24	0.60	0.22
50	1.60	-0.33	0.19
53	1.39	-0.37	0.22
57	1.36	0.22	0.21
58	1.98	0.51	0.23
60	1.41	0.48	0.17
61	1.99	.34	.20
62	1.16	1.20	0.18
63	1.12	0.37	0.23
66	1.75	0.90	0.23
67	2.14	0.91	0.23
69	1.75	1.53	.21
73	0.99	0.19	0.23
74	1.98	2.09	0.21
76	1.22	1.13	0.30
79	1.59	0.43	0.16
80	1.04	0.52	0.22
84	1.83	0.36	0.16
85	1.54	0.69	0.19
86	2.24	0.32	0.19
88	1.20	-1.22	0.21
89	0.93	-0.13	0.23
95	0.68	-0.69	0.21
99	0.76	-2.76	0.21
100	0.47	-0.32	0.21
102	1.08	-0.06	0.19
104	1.64	1.15	0.18
105	1.55	-0.85	0.20
107	1.52	0.39	0.17
109	1.81	0.21	0.19
111	1.24	-0.40	0.23
112	1.27	0.37	0.26
115	0.57	-0.01	0.21
116	1.38	-0.96	0.19
120	2.08	0.27	0.16
121	1.51	-0.42	0.19
123	1.27	-0.45	0.23
124	0.89	-0.98	0.21
125	1.26	-0.46	0.21
132	1.48	0.13	0.19
135	1.39	0.02	0.19
145	0.97	-0.32	0.20
148	2.15	.41	.16
153	1.42	0.87	0.18
154	0.45	1.28	0.20
155	1.40	0.32	0.26
157	1.88	1.96	0.18
161	1.04	1.36	0.16
164	1.86	1.46	0.19
167	2.01	1.93	0.14
169	1.31	1.16	0.21
171	0.79	-1.01	0.21
176	1.13	1.25	0.19
179	1.18	0.64	0.22
181	1.22	-0.22	0.21
184	2.04	0.44	0.18
188	1.31	-0.33	0.21
189	1.01	0.76	0.21
190	0.90	1.45	0.19
192	0.58	-0.50	0.21
196	1.35	-0.05	0.22
197	1.37	0.04	0.22
198	1.65	1.20	0.16
201	1.39	1.51	0.20
207	0.77	-0.35	0.22
209	1.47	-0.07	0.23
220	1.55	1.08	0.21
224	2.14	0.20	0.20
226	2.13	0.57	0.16
231	1.42	-0.10	0.21
232	1.62	-0.83	0.21
233	0.90	-0.24	0.21
237	1.26	-0.53	0.21
238	1.03	-1.18	0.21
240	1.09	-1.28	0.20
243	1.35	0.69	0.22
247	1.24	0.31	0.21
248	1.34	0.63	0.21
251	1.21	0.63	0.23
252	1.37	-0.69	0.20
253	1.43	1.63	0.26
257	1.95	0.73	0.23
258	0.79	-0.63	0.21
274	1.46	1.87	0.22
276	0.81	0.09	0.20
277	1.26	1.96	0.22
278	1.21	2.34	0.22
280	1.40	1.76	0.24
281	0.88	1.04	0.19
285	1.45	0.81	0.20
296	1.76	1.53	0.18
297	1.72	-0.05	0.18
300	1.02	0.74	0.20
307	1.95	0.00	0.23
308	1.23	1.99	0.24
313	1.29	-1.19	0.21
326	0.96	-0.02	0.22
330	1.47	0.43	0.23
339	0.99	-0.77	0.21
343	1.01	-1.45	0.21
344	0.96	-0.48	0.20
348	1.44	-0.54	0.19
350	0.74	0.69	0.22
358	1.84	1.63	0.28
359	1.26	0.27	0.21
363	1.20	0.96	0.21
364	0.77	0.72	0.20
367	0.67	0.35	0.23
369	2.22	1.07	0.12
372	1.78	-0.66	0.20
374	1.22	1.33	0.20
375	0.58	0.81	0.22
377	1.48	0.97	0.19
381	1.86	0.39	0.17
384	1.43	-0.34	0.20
390	1.54	-0.59	0.20
391	1.19	-1.66	0.21
393	2.17	0.43	0.16
394	1.26	0.48	0.19
403	1.47	0.09	0.21
406	1.62	0.41	0.20
412	1.26	-0.68	0.20
415	0.77	-2.33	0.21
416	1.15	0.41	0.22
421	1.54	-1.51	0.20
422	0.72	-0.93	0.21
424	0.96	-1.21	0.21
427	0.64	0.07	0.21
430	1.32	-0.58	0.20
433	1.25	0.64	0.22
438	1.01	-0.25	0.21
442	1.39	1.13	0.18
447	1.36	0.32	0.31
454	1.38	-0.24	0.22
462	1.30	-0.19	0.20
479	0.93	0.63	0.19
480	1.16	-0.31	0.22
494	1.21	0.56	0.18
498	1.80	1.40	0.13
505	0.73	-0.50	0.21
518	0.92	1.84	0.19
519	0.59	-0.82	0.21
524	1.77	1.89	0.14
535	2.23	0.27	0.20
536	1.58	.52	.21
545	1.69	2.48	0.17
550	1.63	0.32	0.21
554	1.55	1.88	0.19
567	0.95	-0.09	0.23
571	0.95	2.10	0.16
584	1.08	1.65	0.22
590	1.23	-0.61	0.23
591	1.33	-0.85	0.21
593	0.83	0.50	0.22
595	1.43	-1.06	0.20
596	1.24	-0.43	0.21
597	1.07	-0.56	0.20
598	1.45	-0.67	0.19
599	1.45	-0.76	0.21
606	0.34	-2.06	0.21
607	.77	.10	.19
612	0.65	-0.24	0.21
615	1.50	0.16	0.20
617	1.55	-0.49	0.19
621	1.91	.75	.20
623	1.13	-0.97	0.22
625	1.46	-1.32	0.20
628	1.05	3.45	0.12
635	1.24	0.13	0.19
636	2.06	-0.74	0.19
641	1.25	-1.30	0.22
644	0.93	-0.57	0.21
653	1.73	0.54	0.24
661	0.87	-0.55	0.22
662	0.90	-0.52	0.20
663	1.49	0.07	0.21

10. PROPUESTA EN RELACIÓN A LA INCORPORACIÓN DE LOS NUEVOS ÍTEMS (PRE-TEST)

Proponemos también un procedimiento para incrementar progresivamente el tamaño del banco de ítems. Si incorporamos nuevos ítems se reducirán las tasas de exposición, reduciremos el peligro de transmisión de ítems entre evaluados y esperamos mejorar algo la calidad de las estimaciones de nivel de inglés. El banco de ítems de eCAT estaba formado por los ítems de los 5 primeros subtests elaborados (subtests del 2 al 6). Proponemos incorporar los ítems propios del subtest 7, que no son ítems de anclaje, formado por 41 ítems. De ellos tenemos alguna información, la más importante la previsión de dificultad (en deciles) obtenida a partir de los juicios subjetivos que realizaron varios jueces nativos (ver **Tabla 23**).

Tabla 23
Ítems propios del sub-test 7,
según su distribución por dificultad y categoría gramatical.

Subtest 7							
Decil de Dificultad	***A-FOR***	***MORF***	***PRAG***	***LEX***	***SINT***	***C-C***	***Total***
1	605	439,431,434	0	0	0	0	4
2	0	0	638	0	0	244,250,353	4
3	0	25,241	0	90,230	0	0	4
4	0	387,522	0	98,110	0	0	4
5	0	385	0	138,486	540	142	5
6	0	0	639	0	97,488	128	4
7	0	370,460	0	0	0	103,249	4
8	0	383,526	0	361,538	0	0	4
9	0	27,533	0	0	180,362	0	4
10	0	0	0	72,268,380,544	0	0	4
Total	1	14	2	12	5	7	41

Se trata de incorporar progresivamente nuevos ítems a las aplicaciones de eCAT. Dada la distribución de θ de la muestra que ha respondido a eCAT, la propuesta es comenzar con los 4 ítems del decil 5 hasta completar 500 aplicaciones, luego se hará lo mismo con los 4 ítems del decil 6,…y así sucesivamente. Las respuestas a estos ítems no se considerarán para la estimación del nivel de inglés de los evaluados, pero nos permitirá aplicar un diseño de calibración on-line para ir incrementando progresivamente el tamaño del banco de ítems. En el siguiente apartado se evalúa la viabilidad de esa propuesta mediante un estudio de simulación.

11. Estudio de recuperación de parámetros de los ítems pre-test y funcionamiento del programa ICL

Se realizó un pequeño estudio de simulación para anticipar el funcionamiento de la anterior estrategia. Se simularon las respuestas de 10000 sujetos a eCAT (a partir de los parámetros originales) con distribución N (0.5,1). Además de eCAT se simularon las respuestas de todos los sujetos a 13 ítems pretest con parámetros *b* de -1.5 a 1.5 en pasos de 0.25, parámetros *a* de 1.30 y parámetros *c* de 0.20. Se manipularon las siguientes variables en la calibración:

- Tamaño muestral (4 tamaños muestrales: los primeros 500 evaluados, los primeros 1000, los primeros 2000 y la muestra total de 10000 sujetos).
- Ítems pre-test calibrados junto con eCAT (3 subtests distintos: los 5 ítems más fáciles, los 5 de dificultad media y los 5 ítems más difíciles).

A cada submuestra resultante de cruzar las anteriores condiciones, se aplicó el método MWP-MEM descrito por Kim (2006a, 2006b) como el mejor método de calibración asumiendo parámetros fijos para los ítems operativos. En esta ocasión, se utilizaron las distribuciones previas por defecto de BILOG: Distribución Log-normal (0,0.5) para el parámetro *a*, sin distribución previa para *b* y distribución beta (5, 17) para el parámetro *c*[3]. También se probó si el añadir para *b* una distribución previa Normal con media igual a la media *b* del subtest analizado mejoraba la estimación de los parámetros de los ítems (la media de *b* para cada uno de los 3 subtest era -1, 0 y 1, respectivamente). En el contexto de tests fijos, Swaminathan et al. (2003) encuentran que se pueden producir algunas mejoras en la estimación de los parámetros para el modelo logístico de 3 parámetros. Queda por estudiar la generalización de sus resultados al contexto de la calibración on-line.

Los resultados (**Tabla 24**) muestran que el parámetro *b* se recuperan razonablemente bien con muestras de 500 sujetos. Con 1000 sujetos el RMSE de cada parámetro alcanza valores similares al error típico del correspondiente parámetro en el

[3] Esto se hizo ya que si se utilizaban las distribuciones previas originales, las medias de las distribuciones previas coincidirían exactamente con los parámetros *a* y *c* de los ítems simulados. De esta forma se podría dar el resultado conceptualmente paradójico de que, en ausencia de aplicaciones, los parámetros serían perfectamente estimados.

diseño de anclaje original (ver Tabla 25). No se ven grandes efectos de incluir o no una distribución previa para el parámetro *b.*

Tabla 24

RMSE para los parámetros a, b y c en función del tamaño de la muestra de calibración y el procedimiento de calibración (15 casos por celdilla)

	Con distribuciones previas en a y c (BILOG default)			***Con distribuciones previas en a y c (BILOG default) + Distribución previa para b acorde con la dificultad del subtest***		
N	***RMSE(a)***	***RMSE(b)***	***RMSE(c)***	***RMSE(a)***	***RMSE(b)***	***RMSE(c)***
500	.170	.121	.038	.167	.110	.038
1000	.150	.096	.027	.148	.094	.026
2000	.117	.065	.039	.119	.067	.037
10000	.035	.032	.008	.035	.033	.008

Tabla 25

Media del Error típico de medida de los parámetros de los ítems en la calibración original

N	***Media del Error típico de medida***	***Error típico de medida dadas las distribuciones previas si no hubiera aplicaciones***
a	.147	.365
b	.102	****
c	.021	.087

12. Referencias

Abad, F.J., V. Olea, J., Ponsoda, V., Ximénez, C. y Mazuela, P. (2004). Efecto de las omisiones en la calibración de un test adaptativo informatizado. *Metodología de las Ciencias del Comportamiento.* Suplemento, 1-6.

Ban, J.C., Hanson. B. H., Wang, T., Ti, 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-C., Hanson, B.A., Yi, Q., y Harris, D. J. (2002). Data sparseness and online pretest item calibration/scaling methods in CAT. *Journal of Educational Measurement*, 39, 207-218.

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

Brumfield, T., Burroughs, R. y Luecht, R. (2001). *Optimal calibration sampling designs for the uniform CPA examination*. AERA. Seattle, WA.

Buyske, S. (1998). Optimal design for item calibration in computerized adaptive testing: the 2PL case. En N. Flournoy *et al.*, editors, *New Developments and Applications in Experimental Design*, volume 34 of *Lecture Notes-Monograph Series*, Haywood, C: Institute of Mathematical Statistics.

Do, B.-R., Chuah, S. C., y Drasgow, F. (2004). *Item parameter recovery with adaptive tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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, Computerized adaptive testing: theory and practice. Boston, MA: Kluwe-Nijhoff.

Guo, F. y Wang, L. (2003). *Online calibration and scale stability of a CAT program*. NCME. Chicago: IL.

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*. NCME. Chicago: IL.

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

Hanson, B. A. (2002). *ICL: IRT Command Language.*

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*(1), 3-24.

Hanson, B. y Zeng, L. (1995). *ST [A computer program for IRT scale transformation, Version 1.0]*. Unpublished. American College Testing.

Holman, R. y Berger, M.P.F. (2001). Optimal calibration designs for tests of polytomously scored items described by item response theory models. *Journal of Educational and Behavioral Statistics, 26*, 4, 361-380.

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

Huang, C.-H., Kalohn, J.C., Lin, C.-J. y Spray, J. (2000). *Estimating item parameters from classical indices for item pool development with a computerized classification test*. ACT Research Report Series 2000-4. Iowa City, Iowa.

Kim, S. (2006a). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43* (4), 355-381.

Kim**,** S. (2006b). A study on IRT fixed parameter calibration methods using BILOG**-** MG. *Journal of Educational Evaluation*, 19 (1), 323-342

Krass, I. A. y Williams, B. (2003, April). *Calibrating CAT pools and online pretest items using nonparametric and adjusted marginal maximum likelihood methods*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago IL.

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 3: Item analysis and test scoring with binary logistic models [computer program]. Chicago: Scientific Software, Inc.

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, 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*, vol. especial, 427-430.

Partchev, I. y Steyer, R. (2004, October). *Scale shift in the online calibration of pretest items: should we fix or equate anything at all?* 46th Paper presented at the annual Conference of the International Military Testing Association, Brussels, Belgium.

Pommerich, M. y Segall, D. O. (2003, April). *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.* Suplemento. 507-510.

Segall, D. O. (2003, April). *Calibrating CAT pools and online pretest items using MCMC methods.* Presented at the Annual Meeting of the National Council on Measurement in Education, Chicago IL.

Stocking, M. L. (1988). Scale drift in on-line calibration (ETS Research Rep. No. 88-28). Princeton, NJ: ETS.

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

Swaminathan, H., Hambleton, R.K., Sireci, S.G., Xing, D. y Rizavi, S.M. (2003). Small sample estimation in dichotomous item response models: effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement*, 27, 27-31.

Raju, NS. *DFIT5P: A Fortran program for calculating dichotomous DIF/DTF [computer program]*. Chicago, IL: Illinois Institute of Technology; 1999.

Wainer, H., y Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. En Wainer, H. (Ed.), *Computer adaptive testing: A primer* (Chapter 4, pp. 65- 102). Hillsdale, NJ: Lawrence Erlbaum.

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.

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

Zwick, R. (2000). *The assessment of differential item functioning in computer adaptive tests.* In W. J. van der Linden C. A. W. Glas (Eds.), Computerized adaptive testing: Theory and practice (pp. 221–244). Boston : 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.

Anexo 1: Distribución conjunta dificultad x categoría gramatical para cada subtest

Test de anclaje							
	A-FOR (1)	**MORF (7)**	**PRAG (1)**	**LEX (5)**	**SINT (3)**	**C-C (3)**	Total
1	1	0	1	0	0	0	2
2	0	2	0	0	0	0	2
3	0	1	0	0	0	1	2
4	0	1	0	1	0	0	2
5	0	2	0	0	0	0	2
6	0	1	0	0	1	0	2
7	0	0	0	1	1	0	2
8	0	0	0	0	1	1	2
9	0	0	0	2	0	1	3
10	0	0	0	1	0	0	1
Total	1	7	1	5	3	3	20

Test aplicados

Subtest 2							
1	1	2	0	1	0	0	4
2	0	4	0	0	0	0	4
3	0	2	1	0	0	1	4
4	0	1	1	2	0	1	5
5	0	4	0	0	0	0	4
6	0	1	0	2	1	0	4
7	0	0	0	1	1	2	4
8	0	0	0	0	2	2	4
9	0	0	0	2	1	1	4
10	0	0	0	4	0	0	4
Total	1	14	2	12	5	7	41

Sutbest 3							
1	0	3	0	0	0	1	4
2	1	3	0	0	0	0	4
3	0	2	1	0	0	2	5
4	0	1	0	0	2	1	4
5	0	3	0	0	0	1	4
6	0	1	0	2	1	0	4
7	0	1	0	2	0	1	4
8	0	0	0	2	2	0	4
9	0	0	0	2	0	2	4
10	0	0	0	4	0	0	4
Total	1	14	1	12	5	8	41

Subtest 4							
1	0	2	1	0	0	1	4
2	0	2	0	2	0	0	4
3	0	2	0	0	1	2	5
4	0	2	0	0	1	1	4
5	0	2	0	0	0	2	4
6	0	2	0	0	1	1	4
7	0	3	0	1	0	0	4
8	1	0	0	1	2	0	4
9	0	0	0	4	0	0	4
10	0	0	0	4	0	0	4
Total	1	15	1	12	5	7	41

Subtest 5							
1	0	2	0	1	0	1	4
2	0	2	1	1	0	0	4
3	0	2	0	2	0	0	4
4	0	2	0	0	1	1	4
5	0	2	0	2	0	0	4
6	0	2	0	0	1	1	4
7	0	2	0	0	2	0	4
8	0	1	0	0	1	2	4
9	1	0	0	2	0	2	5
10	0	0	0	4	0	0	4
Total	1	15	1	12	5	7	41

Subtest 6							
1	0	2	0	1	0	1	4
2	0	3	1	0	0	0	4
3	0	2	0	0	1	1	4
4	0	1	1	1	0	1	4
5	1	2	0	0	1	0	4
6	0	2	0	0	0	2	4
7	0	1	0	2	1	0	4
8	0	0	0	2	1	1	4
9	0	1	0	2	1	1	5
10	0	0	0	4	0	0	4
Total	1	14	2	12	5	7	41

Test por aplicar

Subtest 7							
1	1	3	0	0	0	0	4
2	0	0	1	0	0	3	4
3	0	2	0	2	0	0	4
4	0	2	0	2	0	0	4
5	0	1	0	2	1	1	5
6	0	0	1	0	2	1	4
7	0	2	0	0	0	2	4
8	0	2	0	2	0	0	4
9	0	2	0	0	2	0	4
10	0	0	0	4	0	0	4
Total	1	14	2	12	5	7	41

Subtest 8							
1	0	0	1	1	0	2	4
2	0	3	0	1	0	0	4
3	0	1	0	2	0	1	4
4	1	2	0	0	1	0	4
5	0	1	0	1	1	1	4
6	0	3	0	1	0	0	4
7	0	2	0	2	0	0	4
8	0	0	0	0	2	1	4
9	0	2	0	2	1	0	5
10	0	0	0	2	0	2	4
Total	1	14	1	12	5	7	41

Subtest 9							
1	0	3	0	0	1	0	4
2	0	2	0	1	1	0	4
3	0	2	0	0	1	1	4
4	0	0	0	2	0	2	4
5	0	3	0	0	1	1	5
6	0	1	0	2	0	0	4
7	1	0	0	1	1	1	4
8	0	2	0	2	0	0	4
9	0	1	0	2	0	1	4
10	0	0	1	2	0	1	4
Total	1	14	1	12	5	7	41

Subtest 10							
1	0	2	0	1	1	0	4
2	1	2	0	0	1	0	4
3	0	0	0	3	0	1	4
4	0	3	0	0	1	0	4
5	0	1	0	3	0	0	4
6	0	1	0	1	1	1	4
7	0	3	0	1	0	1	5
8	1	1	0	1	0	1	4
9	0	1	0	1	1	1	4
10	0	0	1	1	0	2	4
Total	2	14	1	12	5	7	41

Subtest 11							
1	0	2	0	1	0	1	4
2	0	3	0	1	0	0	4
3	0	2	0	0	1	1	4
4	0	0	0	2	1	1	4
5	0	2	0	0	1	1	4
6	0	2	0	0	1	1	4
7	0	0	1	2	1	0	4
8	0	1	0	2	0	1	4
9	1	1	0	2	0	1	5
10	0	2	0	2	0	0	4
Total	1	15	1	12	5	7	41

Subtest 12							
1	0	0	0	2	1	1	4
2	0	0	0	2	0	0	3
3	0	2	0	2	1	0	5
4	0	2	1	1	0	0	4
5	1	1	0	2	1	0	5
6	0	2	0	1	0	1	4
7	0	2	0	0	1	1	4
8	0	2	0	0	1	1	4
9	0	2	0	0	1	1	4
10	0	2	0	0	0	2	4
Total	1	15	1	10	6	7	41

Subtest 13							
1	1	0	0	1	1	1	4
2	0	1	0	0	1	1	3
3	0	2	0	0	0	2	4
4	0	2	0	1	0	1	4
5	0	3	0	0	0	1	4
6	0	2	0	1	1	0	4
7	0	2	0	1	0	1	5
8	0	1	1	2	0	0	4
9	0	2	0	2	1	0	5
10	0	0	0	2	2	0	4
Total	1	15	1	10	6	7	41

Subtest 14							
1	0	2	1	0	0	1	4
2	0	1	0	1	0	1	3
3	0	2	0	2	1	0	5
4	0	2	0	1	1	0	4
5	0	0	0	2	1	1	5
6	1	1	0	2	0	0	4
7	0	2	0	2	0	0	4
8	0	2	0	0	1	1	4
9	0	2	0	0	1	1	4
10	0	0	1	0	1	2	4
Total	1	14	2	10	6	7	41

Subtest 15							
1	0	2	0	1	1	0	4
2	0	1	0	1	0	0	2
3	0	1	1	1	1	1	5
4	0	1	0	1	1	1	4
5	0	2	0	1	1	2	6
6	0	1	0	1	0	0	2
7	0	1	0	1	2	2	6
8	0	2	0	1	0	1	4
9	0	1	0	1	0	1	4
10	1	2	0	1	0	0	4
Total	1	14	1	10	6	8	41