Psicothema 2004. Vol. 16, nº 3, pp. 519-525
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG


# Un test adaptativo informatizado para evaluar el conocimiento de inglés escrito: diseño y comprobaciones psicométricas

Julio Olea, Francisco José Abad, Vicente Ponsoda y M. Carmen Ximénez
Universidad Autónoma de Madrid

El presente artículo muestra el trabajo desarrollado para diseñar un test adaptativo informatizado que evalúa el nivel de conocimientos de inglés escrito en castellano-parlantes. Se describe el proceso seguido para la elaboración del banco de ítems, la comprobación de sus propiedades psicométricas, el ajuste obtenido al modelo logístico de tres parámetros y las principales características del algoritmo adaptativo. Se aportan algunos resultados iniciales sobre la validez de los niveles de rasgo estimados. Se comentan los resultados obtenidos en un estudio de simulación, cuyo objetivo es comprobar algunas propiedades de los niveles de inglés estimados (precisión y sesgo). Los primeros estudios de fiabilidad y validez han dado resultados satisfactorios.

*A computerized adaptive test for the assessment of written English: Design and psychometric properties.* This article describes the development of a computer adaptive test to assess the written English level of Spanish speakers. A short description is given of the steps followed for the elaboration of the item bank, the evaluation of its psychometric properties, the fit to the 3-parameter logistic model and the main characteristics of the adaptive algorithm. Some initial results on the validity of estimated ability measures are provided. The article includes the results of a simulation study which aims to obtain information on precision and bias of English level estimates. First results on reliability and validity are encouraging.

En países como Estados Unidos y Holanda, resulta cada vez más familiar la aplicación de tests adaptativos informatizados (TAIs) en contextos de evaluación psicológica y educativa. Pruebas tradicionales de tan amplia aplicación en estos países como el Test of English as a Foreign Language (TOEFL), el Graduate Record Exam (GRE) o el Armed Services Vocational Aptitude Battery (ASVAB) disponen en la actualidad de versiones adaptativas informatizadas. Exámenes de licenciatura, certificación, acreditación o admisión se realizan ya de forma usual mediante TAIs. Drasgow y Olson-Buchanan (1999) y Rojas (2001) exponen los desarrollos iniciales y las mejoras incorporadas a los principales TAIs operativos.

Propuestos inicialmente por Lord (1970, 1980) y puestos a prueba por primera vez por Weiss (1974), la idea fundamental de un TAI es seleccionar de forma dinámica, mediante un algoritmo aplicado en un ordenador, los ítems más apropiados para cada sujeto, según el nivel que progresivamente va manifestando en la prueba. A partir de un banco de ítems calibrado desde alguno de los modelos de la Teoría de la Respuesta al Ítem (TRI), la mayoría de los TAIs proceden mediante una estrategia de ramificación variable para la selección progresiva de los ítems, que requiere establecer: a) un procedimiento de arranque, a partir del cual se determina el primer ítem a presentar, b) un método estadístico (bayesiano o máximo-verosímil) para estimar el nivel de rasgo provisional (y la precisión asociada a tal estimación) de un sujeto tras cada una de sus respuestas, c) un procedimiento para seleccionar, tras una estimación provisional de rasgo, el siguiente ítem a presentar, y d) un criterio para dar por finalizada la prueba. Las diversas opciones que se pueden elegir para establecer estos requerimientos, así como la conveniencia de cada una para objetivos de evaluación diferentes, sus ventajas e inconvenientes, pueden consultarse en Olea y Ponsoda (2003).

En comparación con los tests convencionales de longitud fija, mediante un algoritmo adaptativo se consigue una mejor adecuación entre la dificultad de los ítems y el nivel de rasgo del sujeto, y por tanto una estimación precisa de su nivel de rasgo con la presentación de pocos de ítems y en un tiempo de aplicación reducido. Además, dado que diferentes sujetos reciben ítems distintos, los TAIs garantizan más que los tests fijos que los ítems no sean conocidos antes de la realización del test. Estos beneficios resultan especialmente importantes para los responsables de programas de evaluación a gran escala, donde es necesario aplicar los tests de forma continua a muestras numerosas de personas.

El desarrollo de un TAI hasta que se encuentra operativo es un proceso laborioso y exige conocimientos y destrezas técnicas importantes, así como una importante inversión de trabajo por parte de expertos en Psicometría, en el contenido sustantivo de la prueba y en Informática (mucho más si, como es nuestro caso, el test va a aplicarse a través de Internet). En primer lugar, una vez elaborado el banco de ítems con la colaboración de expertos en el te-

Fecha recepción: 15-10-03 • Fecha aceptación: 27-1-04
Correspondencia: Julio Olea
Facultad de Psicología
Universidad Autónoma de Madrid
28049 Madrid (Spain)
E-mail: julio.olea@uam.es

ma objeto de evaluación, la aplicación de la TRI para la calibración del banco de ítems exige obtener respuestas de muestras numerosas de evaluados en los diferentes subtests establecidos. En segundo lugar, debe programarse un conjunto de algoritmos para la selección progresiva de los ítems, para la estimación de los niveles de rasgo y para el cálculo de la precisión obtenida en cada momento. En tercer lugar, el TAI debe someterse a las oportunas pruebas (empíricas y mediante simulación) para garantizar las propiedades deseables de las estimaciones, así como para aportar datos sobre su precisión y validez. Finalmente, su aplicación a través de Internet, si fuera el caso, exige un trabajo adicional de programación para preservar la seguridad del banco de ítems y para realizar el proceso de *presentación-selección de ítems-estimación de rasgo* en un tiempo imperceptible para el evaluando. De estas consideraciones se sigue que no necesariamente un TAI es siempre la mejor solución. Es preciso analizar bien bajo qué circunstancias y objetivos de evaluación resultan auténticamente eficaces (y rentables) este tipo de pruebas. Wainer (2000) realiza una interesante reflexión sobre sus posibilidades y limitaciones.

El objetivo del presente artículo es mostrar el trabajo desarrollado para diseñar un TAI que pretende evaluar de manera fiable, válida y eficiente el nivel de conocimientos de inglés escrito. Comenzaremos describiendo el proceso seguido para la elaboración del banco de ítems, la comprobación de sus propiedades psicométricas y el ajuste obtenido al modelo logístico de tres parámetros. Posteriormente, describiremos las opciones elegidas en la programación del algoritmo adaptativo: procedimiento de arranque, método de selección de ítems, método de estimación de los niveles de rasgo, procedimientos aplicados para el control de las tasas de exposición de los ítems y criterios de parada. Finalizaremos mostrando algunas propiedades psicométricas del TAI obtenidas empíricamente (validez de las estimaciones de rasgo) y mediante un estudio de simulación (precisión y sesgo de las estimaciones).

El fin último de la prueba es ordenar a los sujetos según su nivel de dominio del inglés escrito, de modo eficiente, y con las condiciones de aplicación que actualmente nos permiten las nuevas tecnologías. El objetivo de eficiencia lo pretendemos cubrir precisamente con la aplicación adaptativa del banco de ítems.

## Construcción del banco de ítems y análisis de sus propiedades psicométricas

### *Elaboración del banco y diseño de anclaje*

Dos especialistas en filología inglesa, con la colaboración de tres profesores de Psicometría, elaboraron un banco inicial de 635 ítems, cada uno de los cuales consta de una frase donde faltan ciertas palabras (el hueco se marca con un asterisco) y 4 opciones de respuesta, una de las cuales es la correcta. Para ello, se siguió un marco teórico funcional-cognitivo, donde se consideraran aspectos de *competencia en el discurso*, que incluyeran el dominio del idioma en situaciones comunicativas específicas, junto a otros que reflejaran estrictamente la *competencia gramatical* con el idioma. Cada uno de los ítems se clasificó en una de 7 categorías de competencia diferentes, incluyendo cada una varias subcategorías de contenido (hasta un total de 46): aspectos formales, morfología, morfosintaxis, pragmática, léxico, sintaxis y categorías compuestas.

Pensando en la aplicación empírica del banco para su posterior calibración, se estableció un diseño de anclaje donde se tuviera en consideración: a) la dificultad previsible de los ítems, y b) la categoría competencial a la que pertenecían. Para disponer de una primera aproximación a la dificultad de cada ítem, 5 profesores de inglés nativos valoraron inicialmente la dificultad de cada ítem asignándole una puntuación entre 1 y 4. Se obtuvo una medida inicial de dificultad sumando las puntuaciones típicas de los 5 jueces en cada uno de los ítems, que fue clasificada en 10 niveles (deciles) de dificultad. Se decidió elaborar 15 subtests, cada uno formado por 61 ítems, 20 de los cuales forman el test de anclaje (común para todos los subtests) y 41 específicos de cada subtest. Tanto los ítems del test de anclaje como los propios de cada subtest se eligieron de forma que representaran adecuadamente la dificultad del banco y la cantidad de ítems que tenía cada una de las 7 categorías competenciales. Más detalles sobre este proceso y algunas comprobaciones adicionales pueden consultarse en Olea, Abad y Ponsoda (2002).

Con objeto de obtener los primeros datos sobre las propiedades psicométricas del banco y su ajuste a un modelo de TRI, se realizó un estudio piloto con el subtest 1, que se aplicó a una muestra de 435 personas adultas españolas de nivel heterogéneo de dominio del idioma inglés: estudiantes de enseñanza secundaria, estudiantes universitarios, estudiantes de filología inglesa y profesores universitarios. Algunos de los resultados fundamentales de este estudio fueron: a) un coeficiente $\alpha$ de 0.91 para el subtest completo (61 ítems) y de 0.81 para el test de anclaje (20 ítems), b) un buen ajuste de los ítems al modelo logístico de 3 parámetros, c) una correlación de Pearson de 0.75 entre las valoraciones iniciales de dificultad obtenidas a partir del juicio de los expertos y el parámetro *b* de los ítems, d) un coeficiente de determinación ($R^2$ corregido) de 0.40 entre la combinación lineal de varios predictores (variables de formación en el idioma) y los niveles de rasgo estimados mediante TRI.

### *Aplicación del banco de ítems: subtests y participantes*

Cinco subtests (del nº 2 al nº 6, en total 225 ítems) se aplicaron a los estudiantes de primer curso de todas las facultades de la Pontificia Universidad Católica de Chile. Los responsables de esta universidad pretenden con esta aplicación derivar a los estudiantes con menor nivel de dominio del inglés a cursos específicos de formación. Los encargados de la aplicación fueron profesores de cada curso, previamente instruidos de forma oral y escrita sobre el procedimiento. Los subtests se aplicaron en soporte de papel y lápiz, dando un tiempo global de 60 minutos para completar la prueba. En total participaron 3.224 estudiantes, asignados aleatoriamente a uno de los 5 subtests ($n_2$= 665, $n_3$= 660, $n_4$= 645, $n_5$= 636, $n_6$= 618), de tal forma que disponemos de las respuestas de la muestra global a los 20 ítems de anclaje, y de los tamaños muestrales referidos para los 41 ítems propios de cada subtest.

Al acabar la sesión los estudiantes informaron sobre a) el tipo de colegio donde estudiaron la enseñanza media (bilingüe-inglés u otros), b) su nivel autopercibido en el idioma (en lectura, escritura y conversación) y c) su formación (educación primaria y secundaria, academias, familia, estancias en países anglosajones y otros).

### *Análisis psicométrico y estudio de unidimensionalidad*

Se realizaron varios estudios sobre las propiedades psicométricas del test de anclaje y de los diferentes subtests. En los cinco subtests, el número medio de aciertos está comprendido entre 28.4

y 32.2, y la desviación típica lo está entre 13.0 y 14.5. A pesar de la asignación aleatoria de los sujetos a los diferentes subtests, el número medio de aciertos resultó significativamente distinto ($p<0.01$) en los cinco subtests, lo que indica la necesidad de proceder a la equiparación de la métrica de los parámetros de los ítems y de los sujetos, que por otra parte estaba ya prevista en el diseño de anclaje establecido. El menor coeficiente $\alpha$ resultó ser 0.94 y el mayor, 0.96. En el test de anclaje, de 20 ítems frente a los 61 de cada subtest, el valor del coeficiente $\alpha$ resultó menor (0.87). La consistencia interna de los diferentes subtests y del test de anclaje, así como los valores medios obtenidos para las correlaciones biseriales ítem-total, indican una fuerte covariación media entre los ítems que componen cada una de las pruebas. Dado que la longitud de cada uno de los subtests es considerable, queda poco margen para la mejora de su consistencia interna.

Respecto al estudio de la unidimensionalidad, y dados los tamaños muestrales disponibles, se realizaron estudios factoriales exploratorios y confirmatorios para el test de anclaje y estudios factoriales únicamente exploratorios para cada uno de los subtests.

Para el estudio de la unidimensionalidad del test de anclaje se obtuvieron las raíces latentes de la matriz de correlaciones tetracóricas con las comunalidades en la diagonal principal, mediante el método de extracción de mínimos cuadrados generalizados. Se obtuvieron tres raíces latentes con valores superiores a 1 ($\lambda_1$= 9.1, % de varianza= 45.6; $\lambda_2$= 1.2, % de varianza= 6.2; $\lambda_3$= 1.1, % de varianza= 5.3). El cociente entre las dos primeras raíces latentes fue 7.32. Asimismo, bajo la solución unifactorial, únicamente un 1.63% de los residuos fueron superiores a 0.10.

El análisis factorial confirmatorio se llevó a cabo mediante el programa LISREL (versión 8.51) poniéndose a prueba el modelo de un factor. Dadas las características de las variables, se utilizó el método de extracción de factores de mínimos cuadrados ponderados (WLS) que lleva a cabo los análisis a partir de la matriz de covarianzas asintóticas (Muthen, 1984). La solución factorial convergió en un proceso de 9 iteraciones. Todos los parámetros estimados ($\lambda_{ij}$) resultaron significativos ($p<0.05$) con magnitudes que oscilaron entre 0.29 y 0.88. Algunos de los índices de bondad de ajuste obtenidos fueron: $\chi^2_{170}$= 500.52 ($p<0.05$), $\chi^2$/gl= 2.94, GFI= 0.99, RMR= 0.05, NNFI= 0.95 y RMSEA= 0.025 ($p>0.05$). Todos estos indicadores, fundamentalmente el índice NNFI (non-normed fit index), que no depende del tamaño muestral empleado, y el índice RMSEA (root mean square error of approximation), junto a su prueba de significación para la hipótesis nula RMSEA $\leq$ 0.05, nos llevan a concluir que existe un buen ajuste global de los datos al modelo de un factor.

Para el estudio de la unidimensionalidad de los cinco subtests no pudo emplearse el mismo procedimiento que para el test de anclaje, dado que las matrices de correlaciones tetracóricas no resultaron positivas definidas y el tamaño muestral no permite estimar la matriz de covarianzas asintóticas. Alternativamente, se utilizó el programa NOHARM (Fraser, 1988) que estima los parámetros $\lambda_j$ y $\tau_j$ del modelo de factor común y la matriz de covarianzas residual de los ítems (McDonald, 1999). Los valores del índice RMSR (root mean square residual) oscilaron entre 0.0047 y 0.0075 e indican un buen ajuste para el modelo de un factor. Asimismo, también se obtuvieron las raíces latentes para cada uno de los subtests de 61 ítems. En todos ellos se obtuvieron trece raíces latentes con valores mayores que 1, siendo la primera notablemente superior a las restantes. El cociente entre las dos primeras raíces latentes toma valores entre 7.44 y 12.13.

*Ajuste, equiparación y estimación de parámetros*

Para la aplicación del modelo de TRI se eliminaron en primer lugar algunos ítems siguiendo algunos criterios psicométricos clásicos (p.e., ítems con correlaciones biseriales bajas o ítems en los que escoger alguna opción incorrecta correlacionaba positivamente con la puntuación total) y observando la adecuación psicométrica de las funciones de respuesta de las opciones. Estas funciones de respuesta de los ítems se obtuvieron mediante el procedimiento no paramétrico de *suavizar con un núcleo* (kernel smoothing) implementado en el programa TestGraf (Ramsay, 1991, 2000).

Siguiendo los criterios clásicos de forma estricta, convenía eliminar 51 ítems (4 del test de anclaje). Sin embargo, se comprobó que en algunos casos se trataba de ítems difíciles en los que, lógicamente, la función de respuesta de la opción correcta era creciente sólo para un intervalo estrecho de habilidad (los sujetos de muy alta habilidad) y/o la función de respuesta de la opción incorrecta era creciente pero posiblemente sólo para el intervalo de habilidad analizado (siendo posiblemente rechazada por sujetos con mayor nivel de habilidad). Por ello, y considerando la posibilidad futura de implementar un procedimiento politómico de puntuación, 41 de estos ítems (todos los del test de anclaje) se mantuvieron para el estudio del ajuste mediante la TRI.

En segundo lugar, se calibraron los ítems según el modelo logístico de 3 parámetros (métrica normal). Para calibrar en la misma métrica los ítems de todos los subtests se utilizó el diseño de calibración concurrente, en el que las respuestas a ítems no aplicados a los sujetos se consideran como *datos perdidos*. En estudios de simulación realizados con un diseño de anclaje similar al del presente trabajo, la calibración concurrente ha mostrado un rendimiento similar a la calibración separada con equiparación posterior (Hanson y Béguin, 2002). Los parámetros fueron estimados por el procedimiento máximo-verosímil marginal bayesiano implementado en el programa 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 el recíproco del número de alternativas y desviación típica 0.025). Abad, Olea, Ponsoda, Ximénez y Mazuela (enviado) muestran la importancia de elegir bien las distribuciones a priori.

Al analizar el ajuste de los ítems al modelo de 3 parámetros, se encontró que 18 ítems (10 de los cuales habían resultado ya problemáticos siguiendo los criterios clásicos) se mostraron desajustados tomando como criterio para el desajuste su valor $\chi^2$ ($p<.01$) acompañado de residuos grandes para algunos niveles de habilidad y/o funciones de respuesta empíricas no monótono-crecientes. Se mantuvieron todos los ítems de anclaje. Por lo tanto, el banco final se compone de 197 ítems.

La mayor parte de los valores del parámetro de discriminación *a* se encontraron entre 0.83 y 1.90 (media= 1.30; desviación típica= 0.32). Para el parámetro de dificultad *b*, el 90% de los valores son medio-altos y se encuentran entre –1.26 y 2.16 (media= 0.23; desviación típica= 1.00). Para el parámetro de pseudo-azar *c*, la distribución se hallaba centrada en torno al valor 0.20 (media= 0.21; desviación típica= 0.02) con la mayor parte de los valores entre 0.16 y 0.25; este valor refleja la calidad de las opciones in-

correctas, puesto que es inferior a 1/4 (recordemos que los ítems tienen 4 opciones incorrectas). La única correlación significativa, con nivel de significación del 1%, ocurrió entre los parámetros *a* y *c* (-0.369), lo que implica que los ítems más discriminativos son más difíciles para los sujetos con bajo nivel de habilidad.

*Función de información*

Una de las herramientas más importantes para caracterizar un banco de ítems es su función de información. La función de información impone una cota a la máxima precisión que puede obtenerse mediante el TAI. En la figura 1 se muestra la función de información para el presente banco. Para el rango de habilidad entre –1.5 y 3 el error típico de medida alcanzable si se aplicaran los 197 ítems del banco está por debajo de 0.2. El banco de ítems funciona mejor para niveles de habilidad medio-altos. Claramente, los niveles de habilidad por debajo de –2.5 no pueden ser estimados con precisión (errores típicos mayores que 0.5).

## Algoritmo adaptativo

El banco final está por tanto formado por 197 ítems y sus correspondientes parámetros estimados (*a*, *b* y *c*). Mediante C++ Builder se diseñó un algoritmo para la presentación-selección sucesiva de ítems, que tiene las siguientes características:

*Procedimiento de arranque:* para comenzar la prueba, se elige un nivel de rasgo de una distribución normal truncada entre -1 y +1, aplicando como primer ítem el que resulta más informativo para dicho nivel. Es, por tanto, un procedimiento de arranque aleatorio entre niveles medios de rasgo, algo usual cuando se piensa aplicar el test en contextos en los que no se tiene información previa sobre el nivel de rasgo de los evaluados.

*Estimación de los niveles de rasgo:* en el algoritmo se incluye un procedimiento de estimación de máxima verosimilitud, empleando el método de aproximación numérica de Newton-Raphson. Como es conocido, mientras se produce un patrón constante de respuestas (todo aciertos o todo errores) no es posible realizar esta estimación. Mientras ocurre esto, el programa asigna una θ provisional obteniendo el punto medio entre el último nivel de rasgo estimado y 2 (si se ha dado un acierto) o -2 (si se ha fallado el ítem). Este procedimiento es una variante del propuesto por Dodd (1990). En el momento en que aparece variabilidad en las respuestas comienza a aplicarse el método máximo-verosímil. Para cada nivel de rasgo estimado se obtiene el error típico asociado; es decir, el valor inverso de la raíz cuadrada de la información que aportan para el último nivel de rasgo estimado los ítems presentados hasta ese momento.

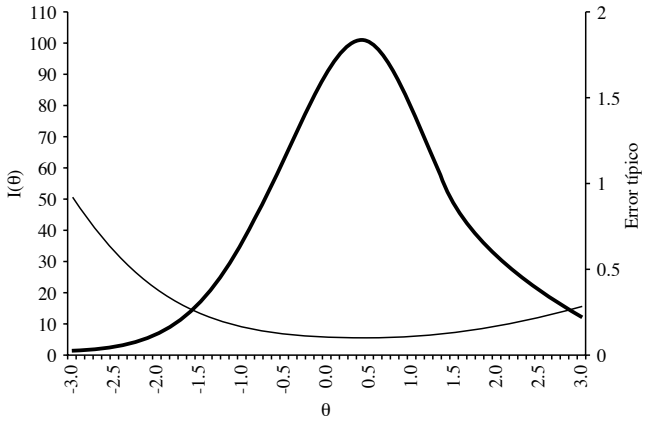


***Figura 1.*** *Función de información y error típico de estimación para el banco de 197 ítems*

*Selección de ítems:* tras la estimación (o asignación) de un valor θ provisional, el algoritmo elige como siguiente ítem, entre los que no se han presentado todavía al sujeto, el que resulta más informativo para dicho nivel de rasgo. En el algoritmo se aplica, por tanto, el método de selección de ítems de máxima información.

Para el control de la exposición de los ítems, se establecen tres restricciones: a) en los 5 primeros ítems que se presentan no son aplicables ítems con parámetros *a* mayores que 1, en línea con lo propuesto por Chang y Ying (1999); b) también en los 5 primeros ítems que se presentan, se aplica el método de McBride y Martin (1983), que consiste en seleccionar como primer ítem uno al azar entre los 5 más informativos; como segundo, otro al azar entre los 4 más informativos;…, y así hasta el 5º, a partir del cuál se selecciona siempre el más informativo para el último nivel θ estimado; c) para todos los ítems se establece una tasa máxima de exposición del 25%, de modo que un ítem deja temporalmente de presentarse cuando ha sido aplicado al 25 % de los sujetos. Con las restricciones a y b vistas se intenta controlar la tasa de exposición de ítems muy discriminativos en las fases iniciales del test, cuando los valores estimados de θ pueden alejarse bastante del nivel verdadero del evaluado; sin esta primera restricción, podríamos aplicar ineficazmente ítems muy informativos para niveles de rasgo alejados del que tiene realmente el evaluado. Con la tercera restricción pretendemos limitar la tasa máxima de exposición al 25% e incrementar, de paso, las tasas de los ítems no demasiado utilizados, En algún TAI hasta el 80 % de los ítems del banco no se aplican nunca o casi nunca (Hornke, 2000), lo que afecta a la seguridad del banco.

*Procedimiento de parada:* el programa permite al responsable de la aplicación establecer diferentes criterios de parada: criterio fijo (estableciendo un número prefijado de ítems para todos los sujetos, con lo cual las diferentes estimaciones tendrán distinta precisión), variable (prefijando un nivel de error típico para todos los sujetos, de tal manera que la presentación de ítems finaliza cuando el error desciende del valor preasignado), o mixto (combinando ambos criterios simultáneamente; es decir, parar la aplicación cuando se presentan *k* ítems o cuando el error típico desciende del valor preestablecido).

## Propiedades psicométricas

*Precisión*

Para estudiar las propiedades psicométricas del test se realizó un estudio de simulación. Se simularon 10.000 sujetos de una distribución normal discretizada en 17 niveles de habilidad entre -4 y 4. Para cada sujeto se simulaban 3 tests adaptativos de longitud fija (30 ítems): a) sin tasa máxima de exposición, b) fijando la tasa máxima de exposición al 40% de las aplicaciones, y c) fijando la tasa máxima de exposición al 25%. En los tres casos se mantuvieron las restricciones *a* y *b* comentadas anteriormente.

En la tabla 1 se muestran las tasas de exposición de los ítems. Si la aplicación de los ítems fuera aleatoria el valor esperado para las tasas de exposición sería del 15%. Puede observarse que, si no se fija una tasa máxima, el 27 % de los ítems son aplicados al menos la cuarta parte de las veces y el 6% más del 40% de la veces.

Imponiendo una tasa máxima de exposición del 40%, el 27% anterior aumenta ligeramente al 28%. En ambas condiciones más del 35% de los ítems se aplican menos del 5 % de las veces. Estos datos muestran claramente la necesidad de imponer restricciones mayores en la exposición. Imponiendo una tasa máxima del 25 %, el 47% de los ítems son aplicados entre el 20 y el 25% de las veces y sólo el 27% de los ítems son aplicados menos del 5 % de las veces.

La tabla 2 muestra el porcentaje de personas que cumpliría el criterio de parada (error típico inferior a 0.30) en función del número de ítems aplicado. Puede observarse que aproximadamente el 85 % de los sujetos son evaluados con precisión adecuada con menos de 20 ítems. Sólo 7 % de los sujetos no cumplen ese criterio de parada, independientemente de la imposición o no de tasa máxima de exposición.

La tabla 2 también muestra el sesgo y RMSE obtenidos. El sesgo observado es pequeño. A partir de 20 ítems se alcanza, en promedio, una precisión adecuada (RMSE <0.30). El coeficiente de fiabilidad (correlación al cuadrado entre el nivel de habilidad real y el estimado) es aceptable incluso con 15 ítems. Por otro lado, el efecto de limitar la tasa máxima de exposición de los ítems apenas deteriora los niveles de precisión.

Una de las ventajas principales de la TRI es que nos proporciona el nivel de precisión obtenido por cada nivel de rasgo. Hemos omitido esta información en la tabla 2, pero resultó evidente que los resultados globales expuestos no son extrapolables fuera del rango de habilidad entre -1 y 2.

*Validez*

Las primeras pruebas de validez se realizaron a partir de los datos obtenidos en el cuestionario comentado en la Introducción. Se realizaron 5 ANOVAs siendo la variable dependiente el valor θ estimado para cada estudiante a partir de sus respuestas al subtest correspondiente (las cinco Fs encontradas resultaron significativas, p<0.001): a) con la variable independiente *tipo de colegio*, los niveles de rasgo medios fueron 0.50 (colegio bilingüe-inglés) y -0.24 (otros colegios). El tamaño del efecto ($\eta^2$) fue 0.10. b) con la variable independiente *formación*, los niveles de rasgo medio fueron -0.16 (sólo colegio), 0.24 (colegio+academia), 0.57 (colegio+familia) y 1.18 (colegio+extranjero). El tamaño del efecto fue 0.09. c) con la variable independiente *autoevaluación de la lectura*, los niveles de rasgo medio fueron -1.16 (nada), -0.86 (sencillo), -0.13 (con esfuerzo), 0.94 (bien) y 1.64 (bilingüe). El tamaño del efecto fue 0.46. d) con la variable independiente *autoevaluación de la escritura*, los niveles de rasgo medio fueron -1.30 (nada), -0.64 (sencillo), 0.03 (con esfuerzo), 0.90 (bien) y 1.77 (bilingüe). El tamaño del efecto fue 0.49. e) con la

*Tabla 1*
Tasas de exposición de los ítems para las condiciones sin control de la tasa máxima de exposición y con control (25% y 40%). Para cada condición se presenta el número de ítems (f), el porcentaje (%) y el porcentaje acumulado (% ac) con una determinada tasa

| | Sin control de la tasa máxima de exposición | | | Tasa máxima: 40% | | | Tasa máxima: 25 % | | |
|---|---|---|---|---|---|---|---|---|---|
| **Tasa** | **f** | **%** | **% ac** | **f** | **%** | **% ac** | **f** | **%** | **% ac** |
| 0.00-0.05 | 73 | 37.1 | 37.1 | 70 | 35.5 | 35.5 | 54 | 27.4 | 27.4 |
| 0.05-0.10 | 19 | 9.6 | 46.7 | 19 | 9.6 | 45.2 | 19 | 9.6 | 37.1 |
| 0.10-0.15 | 16 | 8.1 | 54.8 | 15 | 7.6 | 52.8 | 11 | 5.6 | 42.6 |
| 0.15-0.20 | 19 | 9.6 | 64.5 | 22 | 11.2 | 64.0 | 20 | 10.2 | 52.8 |
| 0.20-0.25 | 17 | 8.6 | 73.1 | 15 | 7.6 | 71.6 | 93 | 47.2 | 100.0 |
| 0.25-0.30 | 16 | 8.1 | 81.2 | 18 | 9.1 | 80.7 | | | |
| 0.30-0.35 | 11 | 5.6 | 86.8 | 12 | 6.1 | 86.8 | | | |
| 0.35-0.40 | 14 | 7.1 | 93.9 | 26 | 13.2 | 100.0 | | | |
| 0.40-0.45 | 6 | 3.0 | 97.0 | | | | | | |
| 0.45-0.50 | 4 | 2.0 | 99.0 | | | | | | |
| 0.50-0.55 | 1 | .5 | 99.5 | | | | | | |
| 0.55-0.60 | 1 | .5 | 100.0 | | | | | | |

*Tabla 2*
Porcentaje de personas que cumplen el criterio de parada, RMSE, sesgo y coeficiente de fiabilidad según el número de ítems aplicados (15, 20, 25 o 30) y el control sobre la tasa máxima de exposición, en la muestra total de sujetos simulados

| Sin control de la tasa máxima de exposición | | | | Tasa máxima: 40% | | | | Tasa máxima: 25 % | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **15** | **20** | **25** | **30** | **15** | **20** | **25** | **30** | **15** | **20** | **25** | **30** |
| Porcentajes | | | | | | | | | | | |
| 65 | 87 | 91 | 93 | 66 | 88 | 92 | 94 | 52 | 83 | 91 | 93 |
| RMSE | | | | | | | | | | | |
| 0.35 | 0.29 | 0.25 | 0.23 | 0.35 | 0.29 | 0.25 | 0.23 | 0.37 | 0.30 | 0.27 | 0.24 |
| Sesgo | | | | | | | | | | | |
| 0.02 | 0.01 | 0.00 | 0.00 | 0.02 | 0.01 | 0.00 | 0.00 | 0.03 | 0.01 | 0.00 | 0.00 |
| $r^2_{\theta\hat{\theta}}$ | | | | | | | | | | | |
| 0.89 | 0.92 | 0.94 | 0.94 | 0.89 | 0.92 | 0.93 | 0.94 | 0.88 | 0.91 | 0.93 | 0.94 |

variable independiente *autoevaluación de la conversación*, los niveles de rasgo medio fueron -1.23 (nada), -0.66 (sencillo), 0.25 (con esfuerzo), 1.01 (bien) y 1.76 (bilingüe). El tamaño del efecto fue 0.53.

En los cinco análisis se observa que los niveles de rasgo medios se incrementan a medida que lo hacen los niveles de cada una de las variables independientes. Todas las comparaciones múltiples post hoc (estadístico DHS de Tukey) resultaron significativas ($p<0.05$). En los valores de los tamaños del efecto ($\eta^2$) puede observarse un mayor poder predictivo de las autoevaluaciones del nivel de inglés que de las variables relacionadas con la formación en el idioma.

Adicionalmente se puso a prueba mediante AMOS (versión 4.01) un modelo estructural para obtener la capacidad predictiva de las estimaciones de rasgo con relación a una variable latente de nivel informado de inglés, donde tuvieran saturaciones positivas las 5 variables evaluadas en el cuestionario. Algunas medidas de ajuste del modelo fueron: $\chi^2/gl = 4.599$, AGFI = 0.992, RMSEA = 0.037. Las estimaciones de los pesos estandarizados se recogen en la figura 2. Puede comprobarse que la correlación entre las estimaciones de nivel de inglés y el factor latente de nivel informado de inglés es 0.81.

## Conclusiones

Las páginas precedentes detallan los pasos realizados para la obtención de un TAI de inglés escrito: la elaboración del banco, el diseño de anclaje, el ajuste y calibración del banco, y las características del programa que estima el nivel de inglés y selecciona los ítems.

Los primeros TAIs operativos pusieron de manifiesto la necesidad de mecanismos de control de la exposición; pues si no, muchos ítems muestran sobreexposición (son presentados a un porcentaje inadmisiblemente alto de sujetos) o infrautilización (no se presentan casi nunca o nunca, reduciendo en la práctica inadmisiblemente el tamaño del banco). El TAI propuesto ha incorporado el procedimiento de control *restrictivo* de Revuelta y Ponsoda (1998), con una tasa del 25%. Los resultados obtenidos no son del todo satisfactorios, pues el porcentaje de ítems infrautilizados sigue siendo muy alto. García Morín y Revuelta (2003) han comprobado que un método que consigue a la vez controlar la tasa máxima (Sympson-Hetter) y reducir la infrautilización (*progresivo*, de Revuelta y Ponsoda, 1998) resulta mejor que el descrito en estas páginas. El nuevo método se ha probado además en condiciones en las que hay control de contenidos. Es decir, la elección de ítems se hace de forma tal que cada test administrado presenta un número similar de ítems de cada área de contenidos.

La principal ventaja de los TAIs es su eficiencia: consiguen medidas precisas con muchos menos ítems que los tests tradicionales. Nuestro TAI mide con una precisión aceptable con solo 20 ítems a más del 80% de nuestros sujetos simulados. En los casos de niveles de inglés extremos, en especial los muy bajos, necesitamos más ítems para alcanzar la precisión deseada. La precisión obtenida ha resultado similar tanto si se controla la tasa máxima como si no.

Se han realizado unos primeros estudios de validez y hemos comprobado que los ítems del banco se relacionan como cabía esperar con las respuestas a un cuestionario en el que se preguntaba por el tipo de formación recibida para el aprendizaje del inglés. La correlación entre el rasgo latente *nivel informado de inglés*, obtenido a partir de las respuestas al cuestionario, y los niveles de rasgo estimado es de 0.81. Es evidente, no obstante, que hay que hacer más estudios de validación.

Wainer (2000) muestra el crecimiento exponencial que ha tenido el número de TAIs administrado durante la última década, y sus repercusiones económicas, sociales y científicas. Una reflexión interesante del trabajo es que precisamente los TAIs se están aplicando en programas de evaluación en los que no son necesariamente la mejor opción. En nuestro país, sin embargo, el interés por los TAIs, que sepamos, no termina de salir de los recintos universitarios (Rojas, 2001). En este sentido, nuestro TAI es una novedad. En colaboración con el Instituto de Ingeniería del Conocimiento de la Universidad Autónoma de Madrid (Ponsoda, Olea, Abad, Aguado, López y Díaz, 2003), nuestro test, con el nombre eCat, ha sido

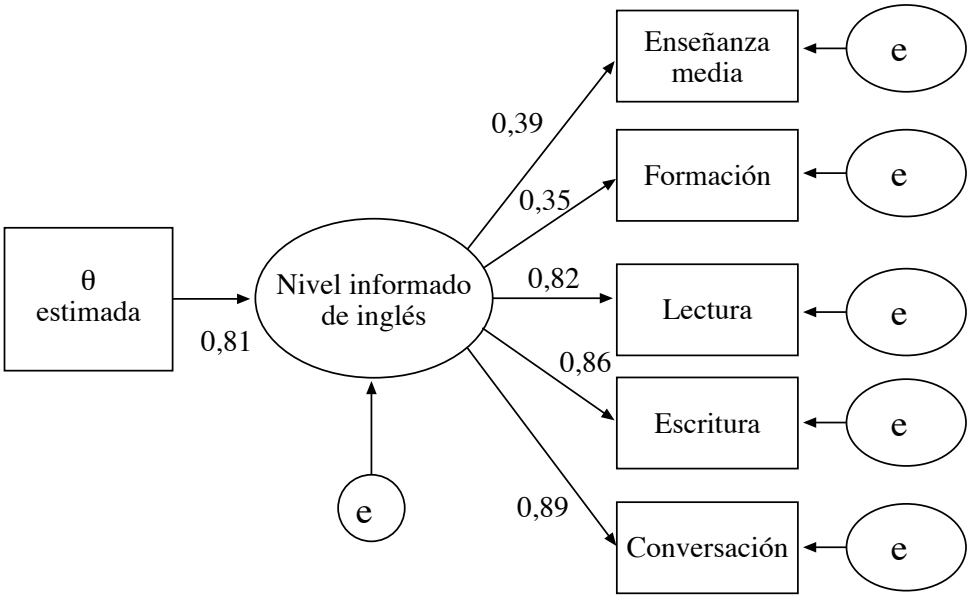


***Figura 2.*** *Modelo estructural para obtener la regresión del nivel informado de inglés sobre la θ estimada*

puesto a disposición de las empresas para la evaluación eficiente del inglés escrito. El test eCaT puede administrarse por internet o instalarse en la red local (http://www.iic.uam.es/flash/eCatiic-flash.html).

Agradecimientos

Este trabajo ha sido financiado por los proyectos DGES BSO2002-1485 y DGES BSO2000-0058.

## Referencias

Abad, F.J., Olea, J., Ponsoda, V., Ximénez, M.C. y Mazuela, P. Efecto de las omisiones en la calibración de un test adaptativo informatizado. *Metodología de las Ciencias del Comportamiento*. Enviado.

Chang, H.H. y Ying, Z. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

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

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

Fraser, C. (1988). *NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. NSW: University of New England.

García Morín, J.R. y Revuelta, J. (2003). Métodos para controlar la sobreexposición e infrautilización de ítems en tests adaptativos informatizados. *VIII Congreso de Metodología de las Ciencias Sociales y de la Salud*. 16 a 19 de septiembre de 2003. Valencia.

Hanson, B.A. y Beguin, A.A. (2002). Obtaining a common scale for IRT item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3-24.

Hornke, L.F. (2000). Item response times in computerized adaptive testing. *Psicológica*, 21, 175-189.

Lord, F.M. (1970). Some test theory for tailored testing. En W.H. Holtzman (Ed.), *Computer assisted instruction, testing and guidance*. (pp. 139-183). New York: Harper and Row.

Lord, F.M. (1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale, NJ: LEA.

McBride, J.R. y Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. En D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223-236). New York: Academic Press.

McDonald, R.P. (1999). *Test theory: a unified treatment*. Mahwah, NJ: LEA.

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.

Muthen, B. (1984). A general structural equation model with dichotomous ordered categorical and continuos latent variables indicators. *Psychometrika*, *49*, 115-132.

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.

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

Ponsoda, V., Olea, J., Abad, F.J., Aguado, D., López, F. y Díaz, J. (2003). eCat. Computerized Adaptive Test para la evaluación del nivel de conocimientos de inglés escrito. VIII Congreso de Metodología de las ciencias Sociales y de la Salud. 16 a 19 de septiembre. Valencia.

Ramsay, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611-630.

Ramsay, J.O. (2000). TestGraf 98: *A program for the graphical analysis of multiple choice test and questionnaire data.* Descargable en la siguiente dirección: http://www.psych.mcgill.ca/faculty/ramsay.html.

Revuelta, J. y Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement* , *35*(4), 311-327.

Rojas, A. (2001). Pasado, presente y futuro de los tests adaptativos informatizados: entrevista con Issac Bejar. *Psicothema*, *13,* 685-690.

Wainer, H. (2000). Computer Adaptive Tests: Whither and whence. *Psicológica*, *21*, 121-133.

Weiss, D.J. (1974). *Strategies of adaptive ability measurement*. Research report 74-5. Dep. Of Psychology, University of Minnesota.