Renom, J., Doval, E. y Sellés, M. (1998). Optimización de los TAI mediante el procedimiento de autoarranque. *RELIEVE*, *4* (2), art. 2. http://www.uv.es/RELIEVE/v4n2/RELIEVEv4n2_2.htm

e-Journal of Educational Research, Assessment and Evaluation

RELIEVE

Revista ELectrónica de Investigación y EValuación Educativa

OPTIMIZACIÓN DE LOS TAI MEDIANTE EL PROCEDIMIENTO DE AUTOARRANQUE

by/por

Renom, Jordi (jrenom@psi.uv.es)
Doval, Eduardo
Sellés, Miguel

Article record
About authors
HTML format

Ficha del artículo
Sobre los autores
Formato HTML

Abstract

The Computerized Adaptive Tests (CAT) have many advantages over the Conventional Test of paper and pencil, but their main disadvantage is that imply to build an item bank (IB) with an important size, and it involves working with high samples of subjects. It all implies that the elaboration process of a TAI is extremely complex and expensive, and it means a brake to the development of this kind of strategies.

In this work we analyse in an empirical way the possibilities of the Self-start procedure making TAIs, a routine to build BI designed by Renom y Martinez (1995) that allow to measure the items partially avoiding the disadvantages pointed out above. Good products are obtained with this method bering in mind the economy of the process and the similarity of the final products in relation to conventional techniques.

Keywords
Computerized Adaptive Tests; CAT; item bank

Resumen

Los Tests Adaptativos Informatizados (TAI) presentan muchas ventajas sobre los Tests Convencionales de lápiz y papel, pero su principal inconveniente radica en que implica construir un Banco de Items (BI) de un tamaño importante, lo que supone trabajar con elevadas muestras de sujetos. Todo ello hace que el proceso de elaboración de un TAI sea sumamente complicado y costoso, lo que supone un freno en el desarrollo de este tipo de medidas.

En el presente trabajo se analizan empíricamente las posibilidades que en la elaboración de los TAI tiene el Procedimiento de Autoarranque, una rutina para la construcción de BI elaborada por Renom y Martínez (1995) que permite calibrar los items del banco evitando parcialmente los inconvenientes apuntados anteriormente. Con este procedimiento se obtienen resultados aceptables teniendo en cuenta la economía del proceso y la similitud de los resultados finales respecto a los obtenidos con técnicas convencionales.

Descriptores
Tests Adaptativos Informatizados; TAI; Banco de Items

1. Introducción

En la década de los años 70, los avances informáticos y los desarrollos en la Teoría de Respuesta al Item (TRI) abrieron las puertas a la posibilidad de realizar unas medidas adaptadas a las características de los sujetos evaluados (McBride, 1997). Desde entonces y hasta el momento, se ha venido subrayado las ventajas que este tipo de Tests Adaptativos Informatizados (TAI) representan respecto a los tests Convencionales (TC) en los ámbitos psicológico y educativo, haciendo hincapié en la capacidad que tienen para realizar evaluaciones de una manera más simple, rápida, dinámica, flexible, precisa y segura.

Si bien es verdad que los TAI ofrecen éstos y otros tipos de ventajas respecto a los TC (ver, por ejemplo, Weiss, 1982, 1983; Wainer, 1990; Renom, 1993; Eignor, 1997), no es menos cierto que los TAI no están exentos de problemas, de manera que éstos ya comienzan a describirse

en la literatura especializada (Bennet, 1994; Craigh y Stocking, 1995; Renom, 1997).

Las ventajas de los TAI sobre los TC se observan principalmente en el momento de aplicar los tests y benefician especialmente a los resultados que se derivan de dicha aplicación. Sus principales inconvenientes, sin embargo, se concentran en la etapa de desarrollo del test, dado que para poder disponer de un TAI es necesario iniciar un proceso de elaboración que resulta muy costoso tanto en tiempo como en esfuerzos invertidos (Craigh y Stocking, 1995).

Debido a esta gran inversión inicial, muchos proyectos e iniciativas dirigidas hacia la construcción tests de este tipo se han visto frustradas, y como consecuencia de ello se ha llegado a la paradójica situación de que en la actualidad, exceptuando algunos pocos países como los EEUU, Holanda o Israel, (Muñiz, 1997) el número de TAI comercializados, y por tanto su uso, es considerablemente inferior a lo esperado si se atiende a sus ventajas tantas veces elogiadas.

En el desarrollo de un TAI cabe distinguir las siguientes etapas (Ver, por ejemplo, Olea y Ponsoda, 1996; Renom, 1993, 1997; Wainer, 1990): creación de items nuevos o adaptación de items existentes, diseño de anclajes de pruebas convencionales de lápiz y papel que compartan items comunes o de ancla, administración convencional de las pruebas, calibración de los items mediante un modelo de la TRI, equiparación de los parámetros de los items respecto a los de anclaje, diseño y edición informatizada de los ítems y, finalmente, elección de un procedimiento de selección y presentación de los items de carácter adaptativo. Así pues, como puede observarse, el grueso del trabajo relacionado con la construcción de un TAI consiste en la elaboración de un banco de items (BI) (Barbero, 1996).

Los principales problemas apuntados anteriormente se deben al hecho de que para construir un buen TAI es necesario disponer de un amplio BI. De acuerdo con Bunderson, Inouye y Olsen, (1998), para poder utilizar un TAI pueda nutrirse de un BI, éste ha de contener al menos 100 items. Estos mismos autores señalan que para calibrar los parámetros de esos items sería necesaria una muestra superior a 500 personas.

Dado la cantidad de items que se precisan, es habitual dividir el BI en bloques o pruebas que contengan un número de items accesible para las personas a las que se vaya a administrar. Este procedimiento implica realizar cuidadosos diseños de anclajes de las pruebas con la finalidad de efectuar una equiparación de los parámetros de los ítems del banco definitivo (Navas, 1996).

El proceso que brevemente se acaba de describir, alarga tediosamente el tiempo transcurrido entre el inicio del desarrollo de un TAI y su aplicación definitiva, y constituye el principal freno con que cuenta el progreso de los TAI en la actualidad (Craigh y Stocking, 1995).

En el presente trabajo se analizan empíricamente las posibilidades que en la elaboración de los TAI tiene el procedimiento de autoarranque (PA), una rutina para la construcción de BI elaborada por Renom y Martínez (1995) e implementada en el programa DEMOTAC2 (Renom y Martínez, 1994), que permite calibrar los items del banco sin necesidad de realizar anclajes ni equiparaciones.

2. Descripción del procedimiento de autoarranque

La calibración de items mediante el PA comienza con la realización de un supuesto sobre las características de los mismos. Se asume que la discriminación es igual para todos los items del banco y que éstos pueden ordenarse según su dificultad. La información apriorística de la dificultad de cada item puede provenir de estudios anteriores, de un peritaje de expertos o simplemente, puede realizarse algún supuesto que no resulte ilógico, como por ejemplo, si se da el caso, el de equidistancia en los valores del parámetro de dificultad. En este sentido también las Redes Neuronales Artificiales constituyen una interesante posibilidad dada su capacidad de predecir la dificultad de nuevos items en base al conocimiento adquirido sobre las respuestas que han recibido otros con característi-

cas semejantes (Perking, Gupta y Tammana, 1995; Renom, Solanas y Sellés, 1997).

De cualquier forma, y con base a esta información inicial y provisional acerca de la dificultad de los items, se realiza una administración adaptativa del BI a un conjunto de sujetos, mediante los procedimientos de Máxima Información, Binivel o Ramificado. A cada sujeto, pues, se le administra un subconjunto de items, y a partir de sus respuestas a ellos se obtiene su nivel de habilidad y el Error Estándar de Medida (EEM) del mismo.

A continuación, se asigna un valor a los items que no han sido presentados al sujeto. La asignación se efectúa de acuerdo con la siguiente regla: si el parámetro de dificultad inicial (*b*) del item tiene un valor inferior al nivel de habilidad estimado para el sujeto menos un EEM, se supone que el sujeto lo acertaría en caso de que se le presentase, y por tanto se le asigna una respuesta con valor 1 (superado). Por el contrario, si el valor *b* del item es superior al de la habilidad estimada más un EEM, se supone que el item resultaría demasiado difícil para ese sujeto, y por tanto se le asigna el valor 0 de respuesta (no superado). Finalmente, a aquellos items cuyas dificultades tienen valores situados alrededor del nivel estimado de habilidad del sujeto ((-EEM) < *b* < (+EEM)) se les asigna de manera aleatoria el valor 0 o 1 como supuesta respuesta del sujeto.

A partir de esta regla arbitraria de asignación de valores de respuesta a los items no contestados por el sujeto, se reconstruye el patrón de respuestas de cada persona. Como algunas de las respuestas de dicho patrón son reales y otras ficticias, podemos denominado Pauta de Respuestas Parcialmente Ficticias (PRPF).

Para finalizar el proceso, a partir de la matriz de PRPF de todos los sujetos se realiza la calibración definitiva de los items y vuelven a estimarse los niveles de habilidad de cada sujeto.

3. Método

Con el fin de valorar las posibilidades del PA en la calibración de un elevado número de items, se han utilizado las respuestas de 800 escolares a 140 de los 180 items que conforman el módulo de sintaxis del Inventario Criterial de Lenguaje (ICL) (ver Renom y Martínez, 1995), una prueba que valora la capacidad de comprensión y realización de oraciones simples, subordinadas y pasivas y el uso correcto de adverbios y proposiciones. La prueba, de administración individual, está compuesta por items de respuesta abierta dicotomizada (acierto-error) con mínima posibilidad de conjetura, de dificultad creciente según la valoración de expertos en la materia.

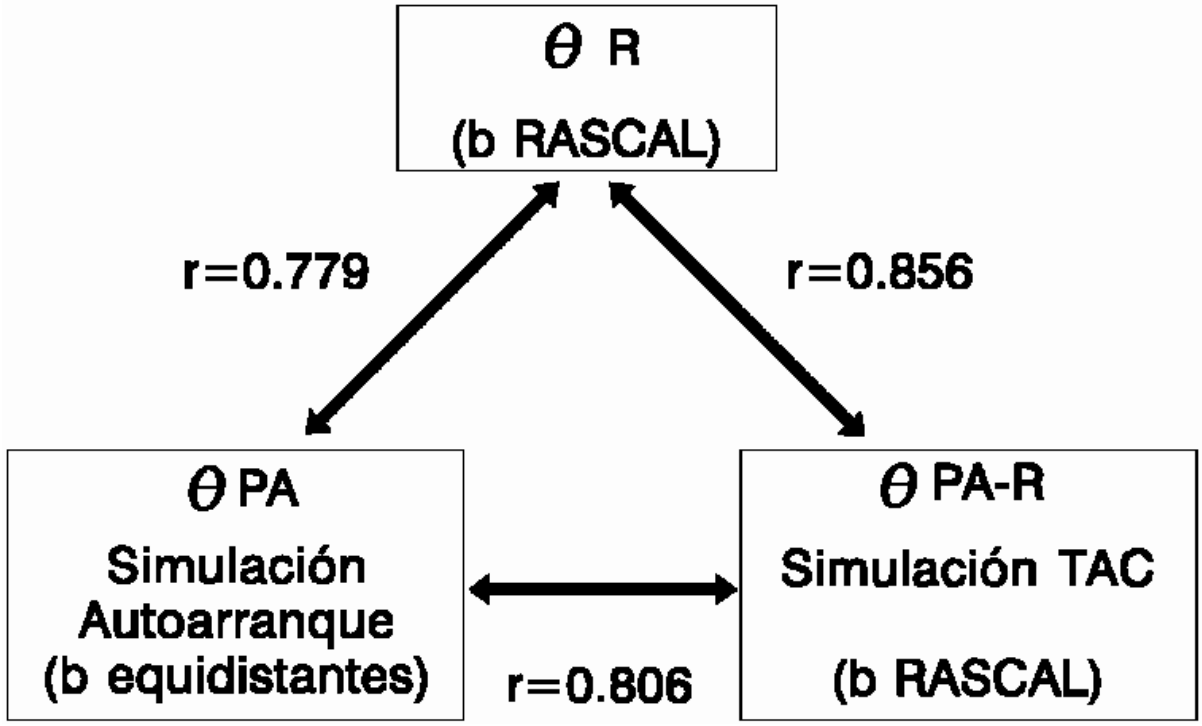
Se han utilizado tres procedimientos distintos para estimar el nivel de habilidad de los sujetos de esta muestra. En primer lugar se ha utilizado el programa RASCAL para ajustar el modelo de Rasch a la matriz de datos y obtener además de una estimación de la habilidad de cada sujeto (R), una calibración de la dificultad de cada item. En segundo lugar, se ha utilizado la dificultad de cada item estimada por el modelo de 1 parámetro, como información inicial en el PA, y a partir de ella se han estimado las habilidades individuales (PA-R). Por último, se ha supuesto equidistancia en la dificultad de los items, previamente ordenados por los expertos, y también empleando el PA se han estimado los parámetros de habilidad de cada sujeto (PA-E). En los dos casos en que se ha utilizado el PA se empleó el procedimiento de Máxima Información para seleccionar los ítems.

4. Resultados

El supuesto de equidistancia en la dificultad de los ítems es realista, dado que guarda una gran relación lineal con los niveles de dificultad estimados mediante el modelo de Rasch ($r=0.9226$, $p<0.001$).

La correlación entre los parámetros de habilidad estimados a través de los tres procedimientos es elevada en todos los casos (ver figura 1). Como era de esperar, es mayor entre R y PA-R (0.856), aunque correlación no difiere considerablemente de la observada entre PA-E y PA-R (0.806) que a su vez es sensiblemente superior a la que muestran las estimaciones de PA-E y R (0.779). Dado el tamaño muestral, todas las correlaciones han resultado significativas al 1 por mil.

Figura 1 - Correlaciones entre las estimaciones del parámetro theta obtenidas mediante el modelo de Rasch (2R), con el PA utilizando las estimaciones de la dificultad de los items mediante el modelo de Rasch (2PA-R) y con el PA suponiendo equidistancia en la dificultad de los ítems (2PA).



El ICl se administra con una finalidad criterial. Se considera que el niño tiene unos buenos conocimientos sintácticos si, como mínimo, contesta de manera adecuada al 70% de los items. Mediante una regresión logística se ha comprobado que con las habilidades estimadas con el modelo de Rasch el nivel criterial real alcanzado por los sujetos puede predecirse correctamente en el 98.88% de los casos, observándose tan solo un 2.08% de falsos positivos y un 0.24% de falsos negativos. Cuando se emplean las habilidades estimadas a partir del PA con las dificultades de los items estimadas mediante el modelo de Rasch, la predicción es del 86.88%, los falsos positivos afectan a un 17.40% de los sujetos que no alcanzaron realmente el criterio, y los falsos negativos al 9.16% de los sujetos que en realidad superaron el criterio. Finalmente, cuando las habilidades de los sujetos se estimaron a partir del PA con dificultades de los items equidistantes, la predicción del nivel criterial de los sujetos fue correcta en el 89.39% de los casos, observándose un 20.51% de falsos positivos y un 1.4% de falsos negativos.

5. Discusión

De los resultados destacan dos hechos principales. En primer lugar que el PA permite realizar unas estimaciones de la habilidad aceptables a partir de supuestos tan básicos sobre los items como el de equidistancia respecto al nivel de dificultad, puesto que las estimaciones de theta realizadas bajo ese supuesto han resultado ser muy similares a las obtenidas, también con autoarranque, a partir de los parámetros de dificultad previamente calibrados con el modelo de un parámetro, y tampoco son muy distintas a las que se han obtenido calibrando los datos con el modelo de Rasch.

En segundo lugar, es evidente que con el PA se ha perdido precisión en las decisiones relativas al comparar la ejecución real de los sujetos respecto al criterio de conseguir responder correctamente a un mínimo de preguntas. Dicha precisión afecta más a la sensibilidad de las estimaciones realizadas, que a la especificidad de las mismas, puesto que con ellas se obtienen más falsos negativos que falsos positivos. En este sentido, se impone un análisis en profundi-

dad de los motivos por los cuales el PA ha dado lugar a este tipo de errores de clasificación. Una posibilidad que creemos muy verosímil, y que pensamos abordar en futuros trabajos, es que en el proceso de creación de las PRPF se generen pautas aberrantes de respuesta.

Podría parecer que la presencia de individuos mal clasificados, valorada de manera absoluta, invalida el uso del PA para la estimación de parámetros. Sin embargo, conviene retrasar dicha valoración hasta tener en cuenta algunos aspectos económicos.

En efecto, utilizando en PA, en el 60% de los casos fueron necesarios menos de 20 items, es decir, un 14% del total de los items, para obtener las estimaciones de sus habilidades, y el número máximo de items necesarios para lograr una estimación de la habilidad fue de 50, o sea, únicamente un 36% del total de los items.

Interpretando estos resultados, puede decirse que el tiempo total de administración del test, que con el método tradicional fue de 2 horas para cada niño, se convierte utilizando el PA en un promedio de 25 minutos. Si tenemos en cuenta que la administración del test es individual y se realiza en horas escolares, hay que sumar 1600 horas aproximadas de trabajo de profesionales que apliquen la prueba, y la misma cantidad de horas de "molestias" a la escuela, cantidad que en ambos casos se reduce a 400 horas si se utilizase el PA. Por último, hay que tener en cuenta que el tiempo real de desarrollo de la prueba fue de aproximadamente medio año, período que se reduciría considerablemente al utilizar el PA, puesto que no sería necesario planificar ni llevar a cabo las fases de anclaje y equiparación que en su momento tuvieron que realizarse (ver Puyuelo, 1994).

Señalar finalmente una ventaja adicional del PA respecto a procedimientos clásicos. Se trata de que con este procedimiento se utiliza desde el primer momento el mismo soporte que luego será utilizado en el TAI, es decir, el ordenador, permitiendo, por ejemplo, iniciar la calibración de items con características multimedia sin necesidad de convertir su formato al de lápiz y papel. Aunque parece que con el paso de un formato a otro no se alteran las principales propiedades psicométricas de los ítems (Hetter, Segall y Bloxon, 1994), sin duda el realizar todo el proceso sin cambiar de formato debe mejorar la validez ecológica del BI.

Referencias

Barbero, I. (1996). Bancos de items. En J. Muñiz (Coordinador). *Psicometría*. Madrid: Universitas.

Bennet, R.E. (1994). *An electronic infrastructure for a future generation of tests*. Comunicación presentada en el Annual Educational Assessment. Wellington, Nueva Zelanda.

Bunderson, C.V., Inouye, D.K. y Olsen, J.B. (1989). The four generations of computerizad educational measurement. En R. L. Linn (Ed.). *Educational measurement*. Londres: Collier Macmillan Publishers.

Craigh, M. y Stocking, M.L. (1995). *Practical issues in large-scale high-stakes Computerized Adaptive Testing*. Research Report-95-23, Educational Testing Service. Princenton, NJ.

Eignor, D.R. (1997). Book review. *Journal of Educational Measurement, 34*(1), 97-100.

Hetter, R.D., Segall, D.O. y Bloxon, B.M. (1994). A comparison of item calibration media in computerizad adaptive testing. *Applied Psychological Measurement, 18*(3), 197-204.

McBride, J.R.(1997). Research antcedents of applied adaptive testing. In W.A.Sands, B.K.Waters y J.R.Mcbride (Eds) Computerized Adaptive Testing: Form Inquiry to operation. Washington, DC: American Psychological Association.

Muñiz, J. (1997). *La medición de lo psicológico*. Lección inaugural del curso académico 1997-1998. Universidad de Oviedo: Servicio de Publicaciones.

Navas, M.J. (1996). Equiparación de puntuaciones. En J. Muñiz (Coordinador). *Psicometría*. Madrid: Universitas.

Olea, J. y Ponsoda, V. (1996). Tests adaptativos informatizados. En J. Muñiz (Coordinador). *Psicometría*. Madrid: Universitas.

Perking, K., Gupta, L. y Tammana, R. (1995). *Predicting item dificulty in a reading comprehension test with an artificial neural network*. Annual testing Research Colloquiu. Illinois.

Puyuelo, M. (1994). *Evaluación de habilidades psicolinguísticas: Proceso de adaptación de una prueba psicométrica*. Tesis Doctoral no publicada. Facultad de Psicología, Universidad Autònoma de Barcelona.

Renom, J. (1993). Tests adaptativos computerizados. Fundamentos y aplicaciones. Barcelona: PPU.

Renom, J. (1997). Retos y perspectivas de los tests adaptativos informatizados. En A. Cordero (Coordinador). *La evaluación psicológica en el año 2000*. Madrid: TEA ediciones.

Renom, J. y Martínez, N. (1994). *LGFI: programa para la creación y administración de items por ordenador*. Comunicación presentada en el III Congreso de Evaluación Psicológica: Santiago de Compostela.

Renom, J. y Martínez, N. (1995). Construcción de Bancos de items sin anclajes ni equiparaciones: el procedimiento de autoarranque. En M. Ato y López-Pina, J.A. (Ed.). *IV Symposio de metodología de las ciencias del comportamiento*. Universidad de Murcia.

Renom, J., Solanas, A. y Sellés, M. (1997). Aplicaciones psicométricas de las Redes Neuronales. Comunicación presentada en el *V Symposio de metodología de las ciencias del comportamiento*.Universidad de Sevilla.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 473-492.

Weiss, D.J. (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. Nueva York: Academic Press.

Wainer, 1990. *Computer adaptive testing: A premier*. Hillsdale, NJ: LEA.

ABOUT THE AUTHORS / SOBRE LOS AUTORES

Renom, Jordi (jrenom@psi.uv.es). Universidad de Barcelona.

Doval, Eduardo. Universidad Autónoma de Barcelona.

Sellés, Miguel. TAD Sistemas.

RELIEVE

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa
E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]