EVALUACIÓN MEDIANTE TESTS: ¿POR QUÉ NO USAR EL ORDENADOR?

Javier López-Cuadrado, Tomás A. Pérez y Ana Jesús Armendariz
Departamento de Lenguajes y Sistemas Informáticos, Universidad del País Vasco, España

1. INTRODUCCIÓN

En muchos sistemas educativos se da la necesidad de disponer de un mecanismo para evaluar la adquisición de conocimientos de los alumnos, algo fundamental para identificar el éxito o fracaso en el proceso de aprendizaje. Los tests son probablemente a día de hoy el modo más habitual de efectuar dicha tarea, pudiendo estar ligados a diferentes contextos educativos. Así, un profesor puede administrar sus propios tests a los alumnos en un momento dado, pero estas pruebas también pueden servir, por ejemplo, para realizar exámenes estandarizados de certificación en un idioma(Hicks, 1989) o incluso estar unidos a un sistema de enseñanza a distancia (López-Cuadrado, Pérez, Arruabarrena, Vadillo y Gutiérrez, 2002).

Las razones para utilizar un ordenador a la hora de evaluar pueden ser variadas y con orígenes distintos, pero todas concluyen en que los tests informatizados ofrecen una serie de ventajas frente a los tradicionales de lápiz y papel(Olea, Ponsoda y Prieto, 1999). En primer lugar, el ordenador es un medio que permite acceder de manera simultánea a datos que se pueden compartir, lo que favorece las condiciones de aplicación de los tests, permitiendo recopilar de manera homogénea y simultánea las respuestas de todos los examinados, requiriendo menos tiempo para ello y reduciendo la posibilidad de copia y de trampa. Además, los ordenadores mejoran el procesamiento de las respuestas, pues dan la posibilidad de ofrecer el resultado de la evaluación en el mismo momento en el que finaliza, el examen, y permiten incluir nuevos tipos de preguntas con elementos dinámicos, multimedia, interactivos o con formatos especiales de respuesta. Por otra parte, los tests administrados en soporte informático proporcionan un mayor control sobre la administración, pues es posible simplificar y configurar los modos de revisión y cambio de respuesta de las preguntas, además de establecer aspectos como el orden en el que se deben responder las cuestiones, el tiempo de presentación de cada una de ellas, y la duración máxima de la aplicación del test. Por último, los tests administrados en soporte informático también evitan el uso de papel y ayudan al desarrollo sostenible del entorno.

A lo largo de este artículo se pretende revisar distintos aspectos a tener en cuenta desde el momento en que se toma la decisión de recurrir a los tests como mecanismo de evaluación, haciendo hincapié en la posibilidad de la utilización de un ordenador para su construcción y administración. Para ello, se partirá de los modelos más sencillos que sustentan su base teórica hasta llegar a los sistemas de evaluación más sofisticados que se plantean hoy en día. Ni qué decir tiene que cuantas más sofisticaciones se deseen incluir en la evaluación mediante tests informatizados, tanto mayor será el esfuerzo necesario para su puesta en marcha.

La siguiente sección está dedicada a presentar las dos grandes teorías psicométricas, centrándose en la más reciente, la Teoría de Respuesta al Ítem (TRI); la sección 3 presenta las fases que se deben seguir durante la construcción de un test de evaluación, desde la concepción hasta la administración del

mismo, tanto si se trata de un test convencional como si está previsto implementar tests adaptativos fundamentados en la TRI a partir de un banco de ítems calibrado; la sección 4 versa precisamente acerca de los Tests Adaptativos Informatizados (TAI); y por último, la sección 5 presenta las conclusiones y el estado actual de la evaluación en los sistemas de e-learning.

2. LA BASE TEÓRICA: LA TEORÍA DE RESPUESTA AL ÍTEM

Una teoría psicométrica proporciona métodos para la construcción de tests y provee de modelos matemáticos que facilitan la interpretación y validación de los resultados obtenidos. Históricamente, la primera teoría fue la que a día de hoy se conoce como **Teoría Clásica de los Tests** (TCT), que surgió a primeros del siglo XX, si bien no recibió su forma axiomática hasta mediados los sesenta (Novick, 1966). La TCT se fundamenta en el denominado modelo clásico, que establece una relación lineal entre la habilidad del examinado y la puntuación obtenida en el test realizado.

Este modelo se expresa en términos de la puntuación empírica del test, el elemento central alrededor del que gira toda la teoría. Concretamente, se considera que la puntuación empírica del sujeto, esto es, el valor observado en el test, es igual a la suma de dos componentes hipotéticos y desconocidos a priori: la puntuación verdadera o habilidad real del evaluando y un determinado error de medida.

La principal limitación achacada a la TCT es que en su contexto las características del test y las del examinado son dependientes, esto es, las mediciones obtenidas dependen por lo general de la naturaleza del test utilizado; y a la inversa, las propiedades de los tests dependen de los sujetos a quienes se les aplica. Así, para la TCT, la habilidad del alumno se mide mediante el número de respuestas acertadas en el test realizado. Por lo tanto, los resultados estarán siempre relacionados con el test administrado: si las preguntas son difíciles, la habilidad de los examinados resultará ser baja porque habrá pocos que las responderán correctamente; y viceversa, si el test es fácil, entonces la habilidad resultante será alta. Por otra parte, la TCT calcula la dificultad de una pregunta (y, por extensión, de un test) en función de la cantidad de individuos que la responden correctamente: cuanto mayor sea el número de evaluandos que responden bien a una cuestión, tanto más fácil se considerará la pregunta. Así, la dificultad estimada de un test según la TCT dependerá de quién lo realice: si los examinados son muy listos, responderán correctamente a las preguntas y, por tanto, el test será considerado fácil. Esta dependencia entre tests y examinados supone que las medidas obtenidas en la TCT no se pueden utilizar en otros contextos de manera generalizada, y por tanto comparar individuos que hayan hecho exámenes distintos resulta harto complicado.

La **Teoría de Respuesta al Ítem** (TRI - Lord, 1952) supera esta limitación, pues propone modelos matemáticos orientados a las preguntas (que en este contexto reciben el nombre de ítems), en contraposición a la TCT, para la que la evaluación de conocimientos gira en torno al test como unidad. Se trata de una teoría relativamente joven, más si se tiene en cuenta que la idea original de los tests adaptativos fundamentados en la TRI, de los que se hablará más adelante, data de comienzos de los años setenta, habiendo sido implementados por vez primera durante los ochenta(Segall y Moreno, 1999).

La gran ventaja de la TRI es que, a diferencia de la TCT, mide la habilidad del examinado y la dificultad de los ítems en la misma escala, lo que facilita su comparación. Se trata del intervalo de valores $(-\infty, +\infty)$, con el valor 0 como punto medio. Si lo consideramos representado en una recta horizontal, los ítems

más fáciles aparecerán a la izquierda del eje y los más difíciles a la derecha. Asimismo, los sujetos con menor habilidad se situarían más a la izquierda que aquellos con mayor destreza.

En la TRI cada ítem se define por una función matemática, la **Curva Característica del Ítem** (CCI), que relaciona la probabilidad de una respuesta correcta y la habilidad del estudiante. En otras palabras, la CCI proporciona para cada ítem la probabilidad P(*è*) de que un alumno, cuya habilidad es *è*, lo responda correctamente. La forma de la CCI la definen varios parámetros que, dependiendo del modelo utilizado, uno puede tener en cuenta o no:

- La **dificultad (parámetro *b*)**, que indica el punto en el que la probabilidad de responder correctamente al ítem es 0.5[1]
- La **discriminación (parámetro *a*)**, que indica la precisión del ítem al determinar la habilidad de los sujetos. Cuanto mayor sea su valor, más significativo resultará el ítem a la hora de evaluar estudiantes, aunque menor será el intervalo de habilidades en el que sea aplicable.
- El **pseudoacierto (parámetro *c*)**, que es la probabilidad de que un examinado con baja habilidad (región izquierda del eje horizontal) responda al ítem correctamente. En otras palabras, puede decirse que se trata de la probabilidad de que el alumno acierte el ítem al azar.
- El **pseudofallo (parámetro *g*)**, la probabilidad de que un examinado con habilidad alta (región derecha del eje horizontal) responda mal al ítem, fallando así un ítem cuya respuesta conoce.

Dado que la TRI se basa en una familia de modelos matemáticos, existen ciertas restricciones o condiciones que se deben considerar sobre los parámetros de los ítems. Así, hay varios supuestos que se deben respetar. Será necesario comprobar las propiedades de **unidimensionalidad** (los ítems sólo miden una única habilidad), **independencia local** (cada ítem es independiente de todos los demás) e **invarianza** (las propiedades de los ítems no dependen de la habilidad de los sujetos, y el nivel estimado del evaluando no se ve afectado por los ítems que componen el test que se le administra).

Un aspecto importante que se deriva de lo anterior es que las curvas características de los ítems incluidos en un test se pueden incorporar por nivel de habilidad definido, obteniendo así la **Curva Característica del Test** (CCT), gracias a la cual es posible predecir, dada la habilidad de un estudiante, cuántas de sus respuestas serán correctas. Asimismo, también será posible realizar estimaciones a priori, por ejemplo, para predecir la probabilidad de que un sujeto con una habilidad específica acierte un determinado ítem.

La TRI promueve el uso de **bancos de ítems**, que almacenan las preguntas susceptibles de formar parte de un test, para desarrollar diferentes pruebas a partir de un mismo banco. Definiendo a priori la curva característica del test que se desea generar, basta con incluir en él aquellos ítems tales que la suma de sus CCI se aproximen a la CCT objetivo. Por otra parte, la TRI también facilita la generación dinámica de tests, en concreto, los Tests Adaptativos Informatizados (TAI) que se tratarán más adelante.

[1] Se ha optado por incluir esta definición de dificultad con un objetivo didáctico. Sin embargo, en la práctica este valor (0.5) variará dependiendo del número de parámetros con los que se trabaje.

3. CONSTRUCCIÓN DE UN TEST DE EVALUACIÓN

Este apartado presenta las fases de construcción de un test de evaluación. La primera etapa es la *concepción del test*, a partir de la que, dependiendo del grado de sofisticación que se desee, podrán o no efectuarse las etapas de *construcción del test* de manera electrónica y *construcción y calibración del banco de ítems*. En todo caso, el desarrollo de la prueba de evaluación culmina en la etapa de *administración del test*, si bien cuando se ha calibrado un banco de ítems conviene realizar periódicamente labores de mantenimiento en las que la *calibración on-line* puede ser de gran utilidad.

3.1 Primera fase: Concepción del test

El primer paso de la administración de un test consiste en crear las preguntas que se pretende incluir, tarea para la que no es necesario en principio ningún ordenador. La construcción de ítems de calidad puede resultar más complicada de lo que parece, de ahí que sea aconsejable seguir algunas sugerencias, como por ejemplo, no incluir en el enunciado palabras como “siempre”, “ninguno” o “generalmente”, no repetir palabras en cada posible respuesta cuando pueden escribirse una sola vez en el enunciado; evitar negaciones dobles y opciones como “ninguna de las anteriores” o “todas las anteriores”; y utilizar tres o cuatro distractores (alternativas de respuesta incorrecta) relevantes y atractivos, con la misma longitud que la respuesta buena (Muñiz, 1997). Aunque seguir estas sugerencias no garantiza el éxito, al menos facilita la identificación de ítems potencialmente problemáticos.

Una vez construidos los ítems del test es posible utilizar los métodos tradicionales de administración del mismo, en concreto, repartir a los alumnos un cuadernillo con las preguntas y una hoja donde inscribir las respuestas a cada una de ellas. En una situación como ésta, en la que puede hacerse caso omiso de las secciones siguientes, la posterior recogida y análisis de datos puede ser una tarea costosa y pesada, al menos si se compara con las facilidades que ofrecen los tests administrados en soporte informático. El siguiente apartado versará acerca de la creación de una versión informática de los ítems desarrollados en esta fase, independientemente de si se les va a dar un uso único en un test o si se pretende almacenarlos en un banco de ítems para aprovecharlos en diferentes pruebas de evaluación.

3.2 Segunda fase: Construcción del banco de ítems o del test

Para poder administrar mediante un ordenador el test diseñado, lo primero que hay que hacer es dar a los ítems un formato adecuado para ser utilizados por el soporte informático. En determinadas situaciones (como cuando los ítems desarrollados no van a reutilizarse) será suficiente con crear el test a administrar mediante un procesador de texto; pero si los objetivos son más ambiciosos (como cuando se quiere tomar la TRI como base para la construcción de tests adaptativos) es necesario implementar un banco de ítems para ser utilizado en la generación de tests. Existen en el mercado múltiples y variadas herramientas que facilitan esta labor, generando automáticamente la base de datos que almacenará el banco de ítems. Aunque algunas de ellas se dedican además a otras tareas complementarias, cabe destacar las aplicaciones Malted (malted.cnice.mecd.es), HotPotatoes 6.0 (www.aula21.net) y My Teacher 2.0, que facilitan la creación de contenidos didácticos y tests de evaluación; así como Test Constructor 2.5, Tester 2.0, TestIt 3.0 Build 110, Random Test Generator PRO 8.0, Academic Test Tool 3.0, QuizMaster 1.0, Examenes 1.2, TestGIP, Aritest Profesores 2.1, y tPilot 1.4, que permiten almacenar ejercicios con el fin de generar tests de evaluación. Todos estos programas se pueden descargar desde la web.

El formato de representación de los ítems puede ser un factor crítico en algunos contextos, de ahí que antes de implementar el banco de ítems sea conveniente decidir cómo se va a simbolizar. Hasta hace poco, cada sistema utilizaba sus propios formatos para representar ítems y tests, tal y como ocurre con los programas recién enumerados. Sin embargo, en la actualidad existe una tendencia a usar estándares para la representación de ítems, como *Question & Test Interoperability* (*QTI*) desarrollado por la iniciativa IMS (www.imsglobal.org). Algunas herramientas como ADISTI (López-Cuadrado, Armendariz y Pérez, 2003) y Canvas Learning (www.imsprojects.org), intuitivas y fáciles de usar, almacenan automáticamente los ítems en una base de datos siguiendo este estándar. Otras herramientas de autor, como Macromedia Authorware 7 y Macromedia Dreamweaver MX con módulo de educación (www.macromedia.com), Tour Virtual de QS Author 1.6 (www.qsmedia.com), o Toolbook 8.6 (www.sumtotalsystems.com), facilitan la informatización de los ítems, permitiendo crear y administrar un curso entero siguiendo algún otro estándar educativo como SCORM (www.adlnet.org) o el propuesto por el AICC (www.aicc.org).

3.3 Tercera fase: Calibración del banco de ítems

Cuando se desea utilizar como marco teórico la TRI, es necesario conocer los valores de los parámetros que definen la curva característica de cada ítem. Aunque la TRI define cuatro parámetros, en la práctica sólo se utilizan los modelos de uno (dificultad), dos (dificultad y discriminación) y tres parámetros (dificultad, discriminación y pseudoacierto).

La calibración consiste en establecer en una métrica común los parámetros de cada ítem del banco. Sólo cuando los ítems se encuentren en la misma escala se podrá asegurar que cualquier subconjunto de ellos proporcionará estimaciones de habilidad invariantes e independientes de la composición del test utilizado. Realizar la calibración de un banco de ítems, si bien no es excesivamente complicado, conlleva tareas largas y costosas, debidas a la gran cantidad de trabajo de campo que se requiere. Una práctica utilizada con cierta frecuencia, aunque al margen de las instrucciones que proporciona la psicometría, es hacer una estimación de la dificultad de cada uno de los ítems (en particular, al utilizar el modelo de un único parámetro) en base a las contribuciones de expertos en la materia que se pretende evaluar. Aunque consultar a profesores o pedagogos doctos en la materia que se pretende evaluar y pedirles que valoren los parámetros de los ítems puede ser un buen comienzo, no es recomendable conformarse sólo con esto, dado que por tratarse de una estimación subjetiva, no siempre resulta fácil determinar acertadamente los valores de los parámetros, y la precisión y validez de los tests posteriormente compilados podría quedar en entredicho. Lo más habitual y recomendable de cara a generar tests adaptativos fiables es calibrar el banco de ítems mediante algún procedimiento estadístico. Por ello, la calibración se ejecuta por lo general en cuatro pasos (Renom y Doval, 1999): primero se administran los ítems a una gran muestra de sujetos, generalmente utilizando algún tipo de *diseño de anclaje*; tras analizar las respuestas recopiladas, se *estiman estadísticamente* los parámetros de los ítems y las habilidades de los sujetos; después se *unifican las escalas* de los diferentes subtests de anclaje para que todo el banco de ítems (y los tests generados a partir de él) utilicen la misma métrica; y por último, se efectúan *estudios de ajuste* de los datos al modelo de la TRI con el fin de identificar y retirar ítems defectuosos. Los siguientes cuatro epígrafes describirán cada una de estas fases.

Diseño de anclaje y administración de los ítems

Los modelos matemáticos de la TRI se fundamentan en variables (parámetros) latentes, difícilmente observables pero que se pueden estimar. Y en esto consiste precisamente la calibración de un banco de ítems. Se trata de administrar las preguntas a una muestra de sujetos, cuyas habilidades son en principio desconocidas, para obtener estimaciones de los parámetros de cada ítem a partir de las respuestas recopiladas. Para poder asegurar que estos parámetros sólo dependen del ítem y no, por ejemplo, de los sujetos a los que se ha administrado, la muestra utilizada ha de ser lo suficientemente grande y heterogénea como para que las estimaciones obtenidas sean insesgadas. Así, el primer paso en el proceso de calibración consiste en administrar cada ítem a una muestra de varios cientos de personas. Llevar a cabo una administración de semejantes características obligará probablemente a repartir los ítems entre diversos subtests. Existen varias alternativas para unificar las previsiblemente diferentes métricas obtenidas en los distintos subtests en una escala que sea común a todo el banco de ítems, pudiendo haber algunas cuestiones que contesten todos los sujetos y/ o algunos sujetos a los que se les administre todo el banco de ítems(Kolen y Brennan, 1995). El objetivo en cualquier caso es disponer de una referencia común a todas las pruebas que sirva de anclaje en la posterior fase de equiparación de las diferentes métricas. La opción más utilizada es la de los ítems de anclaje, que son conjuntos de ítems que dos o más subtests tienen en común. Los parámetros de estos ítems comunes se estiman junto con los del resto de ítems que componen cada subtest, para después comparar los resultados obtenidos en cada caso, lo que facilitará la equiparación de las estimaciones de los parámetros de los ítems no comunes.

Análisis previos y estimación de parámetros

Registrados los resultados de la administración de los ítems, es recomendable realizar análisis previos a la estimación de parámetros con el fin de detectar y depurar anomalías. Renom y Doval (1999) enumeran tres frentes de acción a la hora de analizar las matrices de respuesta: filtrado de la obtención y captura de datos a fin de evitar tratar protocolos anómalos de los examinados, análisis convencionales de cada subtest para detectar ítems incompatibles con los modelos de la TRI, y verificación de las pautas de respuesta de los examinados. Antes de proceder con la estimación de parámetros, también se suele realizar otro estudio, el del supuesto de unidimensionalidad del banco de ítems. Si bien este análisis pertenece a la etapa posterior de verificación del ajuste al modelo de la TRI, su práctica suele adelantarse porque no requiere conocer de antemano los valores de los parámetros. Como resultado de los estudios previos a la estimación de parámetros, puede ocurrir que alguno de los ítems del banco sea retirado del mismo (por ejemplo, por no satisfacer el principio de unidimensionalidad).

Una vez revisadas y depuradas las matrices de respuesta obtenidas tras la aplicación de los subtests se está en condiciones de proceder a la estimación de parámetros en base a alguno de los modelos de la TRI. Cuando se trata de ítems de respuesta múltiple dicotómicos (esto es, en los que sólo se distingue acierto y error), la experiencia y la intuición indican que el modelo de tres parámetros es el más adecuado, algo en lo que coinciden la mayoría de los autores(Santisteban y Alvarado, 2001).

Estimar la habilidad del examinado cuando se dispone de los parámetros de los ítems puede realizarse de manera sencilla mediante la técnica de máxima verosimilitud condicionada (tal y como ocurre en los TAI, de los que se hablará más adelante). Lo mismo ocurre en la situación inversa, esto es, cuando se desea obtener la curva característica de un ítem conocidas las habilidades de los sujetos a quienes se les ha administrado (Baker, 1992). Sin embargo, en el contexto de la calibración del banco de ítems tanto la

habilidad de los sujetos a quienes se les ha administrado los subtests como los parámetros de los ítems son variables desconocidas. Por ello, pese a que sólo interesan las estimaciones de los parámetros de los ítems, es necesaria una estimación simultánea mediante algún método alternativo. La estimación **máximo verosímil conjunta(Birnbaum, 1968)**, que se suele implementar mediante un tratamiento multivariado del procedimiento de Newton-Raphson, asigna un valor inicial (por ejemplo, aleatorio) a los parámetros de los ítems y, asumiendo que son los verdaderos, estima las habilidades de los sujetos, generalmente mediante el procedimiento de máxima verosimilitud condicionada. Tomando estos valores de habilidad recién calculados como reales, se procede a recalcular los parámetros de los ítems (mediante el procedimiento de estimación máximo verosímil, condicionado en este caso a los valores de habilidad). Estas nuevas estimaciones de los parámetros se usarán a su vez para volver a estimar las habilidades de los sujetos, habilidades que permitirán obtener valores más precisos de los parámetros de los ítems. Las dos etapas del proceso se repetirán hasta obtener convergencia en los parámetros de los ítems y las habilidades de los examinados. La estimación conjunta de habilidades y parámetros plantea dos inconvenientes: por una parte, exige la eliminación de las puntuaciones extremas (todo aciertos o todo fallos), tanto para ítems como para sujetos, y por otra, el número de parámetros y habilidades a estimar aumenta a medida que crece el tamaño de la muestra. El método de estimación máximo verosímil marginal(Bock y Aitkin, 1981) evita estos problemas, asumiendo que la muestra de sujetos se ha seleccionado aleatoriamente de una población en la que la habilidad está distribuida en base a una función de densidad g(è), que desde un punto de vista bayesiano correspondería a la distribución previa de probabilidades, en lugar de usar un valor è para cada examinado. A diferencia de la estimación máximo verosímil conjunta, el procedimiento de máxima verosimilitud marginal proporciona consistencia a la estimación de los parámetros, y es independiente del tamaño de la muestra. Pese a ser probablemente la técnica más utilizada, el método de máxima verosimilitud marginal no está exento de problemas, por lo que se han definido algunas variantes y generalizaciones del mismo, habiéndose propuesto incluso alternativas puramente bayesianas (Hambleton y Swaminathan, 1985). Aunque se han presentado las diferentes técnicas de estimación conjunta de parámetros y habilidades, uno puede despreocuparse a la hora de calibrar un banco de ítems, pues existen paquetes de software que las implementan, calculando en pocos segundos estimaciones de los parámetros invariantes y robustas que se ajusten a su curva característica según el modelo TRI correspondiente. Destacan LOGIST (Wingersky, 1983), que implementa la estimaciones máximo verosímil conjunta e incondicional, y se ha convertido en el estándar de facto con el que se comparan los demás procedimientos de estimación de parámetros; y BILOG (Mislevy y Bock, 1990), que se perfila como uno de los mejores programas al implementar la reformulación del método de máxima verosimilitud marginal de Bock y Aitkin (1981).

Equiparación de puntuaciones

Administrar todo el banco de ítems a cada sujeto de la muestra tiene la ventaja de que se elimina una de las fuentes más importantes de error en la equiparación de puntuaciones, a saber, la relativa al muestreo de los examinados. No obstante, plantea diversos problemas, dado que aplicar un elevadísimo número de ítems a una misma persona no siempre es factible, amenaza la seguridad del banco de ítems y puede deparar resultados negativos debidos a la fatiga o a la desmotivación. Por su parte, distribuir los ítems en varios subtests tiene la ventaja de que no se administra todo el banco a cada examinado, pero, después de haber estimado los parámetros de los ítems administrados en los subtests, resulta necesario

equiparar sus escalas de medida con el fin de que todo el banco utilice una métrica común. Sólo así, una vez se dispone del banco calibrado, o lo que es lo mismo, cuando los parámetros de todos los ítems están expresados en la misma métrica, será posible verificar la bondad de ajuste, hecho lo cual se podrá obtener la curva característica y la función de información de cualquier ítem o test generado a partir del banco.

La equiparación de puntuaciones es un proceso estadístico que permite ajustar las puntuaciones de diferentes tests, cuyas dificultades probablemente serán desiguales, con el fin de poder compararlas en una escala de habilidad con origen y unidad comunes. Técnicamente, cuando se ha utilizado un diseño de anclaje para la administración de los ítems, se dirá que éstos están calibrados una vez se haya efectuado la equiparación de sus parámetros (mientras tanto, estarán simplemente estimados) mediante un reescalado lineal de los parámetros de cada subtest a una métrica común. Se han propuesto diferentes métodos para obtener los valores de la pendiente y ordenada en el origen que definen el escalado para cada subtest. Entre las técnicas de equiparación cimentadas en la TRI que permiten expresar las puntuaciones de varios subtests que comparten un diseño de anclaje de ítems, destacan los métodos basados en los momentos (media-sigma, media-media), los métodos basados en la curva característica del test (Haebara, Stocking-Lord, ÷2 mínimo) y el método de la calibración concurrente. La mayor parte del software de estimación de parámetros existente implementa alguno de estos métodos, por lo que uno tampoco debería preocuparse por cuál es el funcionamiento de cada uno de estos procedimientos.

Estudios de ajuste al modelo

Los modelos de la TRI fundamentan su flexibilidad en la realización de suposiciones muy restrictivas que no siempre se ajustan a la realidad. Por este motivo es tan importante este paso, consistente en verificar si las estimaciones recién obtenidas se ajustan al modelo elegido y si se cumplen las restricciones que impone el mismo. La más importante es la comprobación de unidimensionalidad, que consiste en verificar que los ítems sólo sirven para medir una única habilidad. Como ya se ha adelantado, este supuesto puede estudiarse antes de la estimación de parámetros, quedando para después otro tipo de estudios como los de bondad de ajuste de los parámetros de los ítems, los de invarianza de los parámetros, o los de simulación del comportamiento del modelo. Como resultado de esta etapa puede ocurrir que algunos ítems sean retirados del banco por no respetar los supuestos de la TRI.

3.4 Cuarta fase: Administración del test

Una vez se tiene construido el banco de ítems o el test y, en su caso, calibrados sus ítems, se plantea el momento de administrar el o los tests a los sujetos a evaluar. Para este cometido es posible utilizar un método que recoja los resultados a través de un sistema de información, o una aplicación informática que únicamente presente los ítems creados en la segunda fase de la construcción del test. Son muchos los sistemas que automatizan la administración de tests, llegando a presentar características muy diferentes unos de otros. Así, algunos programas como TestGIP, Exam Software 2.3, Aritest Profesores 2.1 y tPilot 1.4, sirven para evaluar al alumno suministrándole un test cuyos ítems tienen almacenados; otras aplicaciones, por su parte, sirven además para mostrar algún tipo de unidad didáctica o lección previa en torno a la cual se desarrollará la evaluación. La ventaja de estos sistemas es su sencillez en la administración, ya que están pensados para que profesores que no están muy familiarizados con la tecnología puedan utilizarlos con facilidad. Sirvan como ejemplo los programas Malted, HotPotatoes 6.0 y My Teacher 2.0, de los que se ha hablado antes. Un tercer tipo de programas informáticos son los que,

además de lo anterior, siguen estándares como SCORM o AICC e incorporan nuevas funcionalidades educativas como la de evaluar y guardar los resultados para un posterior análisis; tal es el caso de sistemas como ELSA (Armendariz, López-Cuadrado, Tapias, Villamañe, Sanz-Lumbier y Sanz-Santamaría, 2003), o las anteriormente mentadas Toolbok 8.6, Tour Virtual de QS Author 1.6 y Macromedia Dreamweaver más el módulo de educación de Macromedia.

Independientemente de qué categoría de software se use, una vez concluido un test de evaluación se dispone de una estimación de la habilidad del examinado. En el marco de la TRI se trata de un valor numérico ($\hat{\theta}$) en la escala de medida del banco de ítems, algo que puede no resultar informativo para el sujeto, de ahí que suela inferirse algún otro tipo de puntuación más significativa. Por ejemplo, este valor puede transformarse a la escala [0,10] o a la métrica de la curva característica del test utilizado, baremarse mediante el uso de centiles o porcentajes acumulados, o incluso representarse gráficamente sobre el continuo de habilidades.

3.5 Quinta fase: Calibración on-line

En el contexto de los tests adaptativos fundamentados en la TRI, de los que se hablará en la siguiente sección, cuando haya pasado algún tiempo desde que se calibrara el banco de ítems, lo más recomendable es disponer de nuevos ítems para añadir, con el fin de sustituir a otros que conviene retirar por haber quedado obsoletos, estar defectuosos o haberse utilizado muy a menudo (Wainer y Mislevy, 1990). El principal problema es que es necesario calibrar estos nuevos ítems en la misma métrica que utiliza el banco. Para realizar esta equiparación existen diversos métodos, uno de los cuales consiste en desarrollar un nuevo proceso de calibración, según se acaba de ver en apartados anteriores. Aunque esta vía puede ser la más adecuada cuando se dispone de muchos ítems nuevos, lo más habitual es que la inclusión sea progresiva y se dé con pequeños conjuntos de ítems, por lo que en este punto será mejor aprovechar que se dispone del banco de ítems calibrado para facilitar el trabajo de cara a estimar los parámetros de los nuevos ítems (López-Cuadrado, Pérez et al., 2002). Así, lo más habitual es administrar a una población numerosa, generalmente la misma a la que se pretende evaluar a partir del banco calibrado, un test compuesto por los ítems nuevos y algunos (de anclaje) pertenecientes al banco. De este modo, el subconjunto de ítems de anclaje permite establecer una conexión entre la métrica de la nueva calibración y la del banco. La denominada *calibración on-line* simplifica aún más el proceso, y en lugar de generar tests específicos, lo que hace es aplicar al comienzo de cada test adaptativo uno o dos ítems, que no influyen en la estimación final de habilidad del sujeto. El objetivo será construir una matriz de datos con la que después realizar la calibración aplicando un sistema de anclaje-equiparación, algo que, a diferencia de lo que ocurre en la calibración inicial del banco de ítems, en este caso resulta muy sencillo porque se dispone de las estimaciones de habilidad de los alumnos (obtenidas por los tests aplicados).

En resumen, en un diseño de calibración *on-line* los nuevos ítems pueden administrarse linealmente junto con los ítems operacionales, para posteriormente ser calibrados y equiparados según la escala del banco actual. Casi todos los programas de evaluación basados en la TRI incluyen la calibración e inclusión de nuevos ítems cada cierto tiempo, principalmente por motivos de seguridad. De hecho, la principal ventaja de la calibración *on-line* se refiere al ahorro temporal y de recursos, pues permite mantener la seguridad de las pruebas sin necesidad de realizar continuos procesos de calibración tan complejos como el descrito en la sección anterior.

4. LOS TESTS ADAPTATIVOS INFORMATIZADOS

Los tests convencionales suelen incluir ítems que abarcan todo el rango de habilidades que se pretende evaluar, de modo que la compilación de este tipo de tests suele consistir en escoger muchas preguntas de dificultad media, y unas pocas de dificultad extremadamente alta y baja. En una situación así, los examinados más hábiles deben responder a un elevado número de ítems que para ellos son fáciles, sin que aporten información valiosa acerca de su nivel de competencia: en casos como éste se sabe que el alumno es muy hábil, pero no se sabe hasta qué punto lo es. Además, como consecuencia directa, se puede producir una considerable desmotivación y aburrimiento por parte del administrado, quien no hace sino responder preguntas (para él) fáciles. En el caso de los evaluandos poco diestros la situación es similar: aunque los ítems fáciles proporcionan cierta información sobre la habilidad del alumno, las preguntas difíciles no sólo no aportan conocimiento al respecto, sino que además pueden causar desconcierto y frustración en el examinado. Para evitar este tipo de situaciones, un examinador experto consideraría que si el evaluando ha respondido mal a una cuestión es porque ésta era difícil para su nivel de habilidad, de manera que la siguiente pregunta que le formule será, en mayor o menor medida, más fácil. Igualmente, ante una respuesta correcta, el examinador inteligente propondrá una pregunta más difícil. La idea subyacente consiste en administrar al alumno únicamente ítems que realmente aportan información útil de cara a medir su habilidad, concretamente ítems cuya dificultad ronda el nivel de destreza que se sospecha tiene el alumno.

Los Tests Adaptativos Informatizados (TAI) son la implementación de este comportamiento inteligente en un programa informático que automáticamente selecciona el ítem más adecuado para administrar en base a las respuestas que haya dado el examinado hasta el momento. Los TAI integran las ventajas psicométricas de la TRI con la capacidad de procesamiento de los ordenadores personales actuales, de modo que hacen posible estimar eficientemente el nivel de habilidad de los alumnos. La finalidad de un TAI es aplicar a cada sujeto únicamente aquellos ítems que realmente aportan información útil acerca de su nivel de habilidad, dado que las preguntas que resultan demasiado fáciles o difíciles no la proporcionan. En relación a este aspecto, la principal ventaja de los TAI frente a los tests tradicionales es la obtención de medidas más precisas habiendo administrado un menor número de ítems. Para cada evaluando se genera dinámicamente un test potencialmente diferente a partir del mismo banco calibrado de ítems, en función de la competencia demostrada durante el desarrollo del mismo. Además, por disponer de un banco de ítems calibrado, será posible equiparar las habilidades de dos sujetos incluso aunque en sus tests no hayan respondido ningún ítem en común.

A las características propias de cualquier test informatizado (enumeradas en la introducción) habría que añadir las ventajas que proporcionan los TAI gracias a su naturaleza adaptativa. En primer lugar, estos tests reducen aún más el tiempo de aplicación, ya que, para obtener una estimación de la habilidad del evaluando con una determinada precisión, los TAI requieren la mitad o un tercio menos de ítems que los tests convencionales. En consecuencia, los efectos negativos que la fatiga del examinado puede generar en los resultados disminuyen considerablemente. Por otra parte, los TAI incrementan la seguridad del test, dado que la mayor parte de los ítems que se presentan a cada evaluando es diferente en cada caso, por lo que difícilmente los alumnos podrán acertar las preguntas por reconocimiento y memoria. Además, los TAI realizan estimaciones más precisas, bajo condiciones similares, que las de un test convencional (en tiempo requerido y en número de ítems utilizados), algo que ya se ha justificado al decir que los TAI escogen en

cada momento el ítem que más información aporta con respecto a la habilidad estimada del evaluando. Dado que la generación de los TAI es dinámica, es posible componer tests bajo demanda sin ningún tipo de coste adicional, lo cual resulta muy ventajoso, por ejemplo, en entornos educativos *on-line* como el sistema hipermedia adaptativo Hezinet, en los que se permite que el alumno decida en qué momento del aprendizaje quiere ser evaluado (López-Cuadrado, Pérez et al., 2002).

Cualquier tipo de test se administra siguiendo un algoritmo de aplicación, es decir, respetando unas reglas que definen cuál es el orden en que se le van a presentar los ítems al evaluando. Aunque existen multitud de algoritmos diferentes, todos ellos tienen una característica común, y es que se definen en base a la respuesta a (1) ¿cuál es el primer ítem que se va a administrar al examinado?, (2) ¿qué ítem se va a administrar después de cada respuesta?, y (3) ¿cuándo se deja de administrar ítems?. El algoritmo para el caso de un test convencional de lápiz y papel es trivialmente sencillo: se comienza por el primer ítem de la hoja, se continúa respondiendo el siguiente ítem de la secuencia, y se finaliza cuando no quedan más preguntas por contestar. De hecho, es en el caso de los TAI cuando la complejidad en los algoritmos de aplicación se hace patente, porque la secuencia de ítems administrados no se conoce a priori, sino que depende de las respuestas que el examinado ha dado a ítems previos. Los siguientes tres epígrafes describirán el modo en que los TAI responden a las tres preguntas aludidas.

1.1. ¿Cómo empezar?

La administración de un test adaptativo comienza una vez se haya seleccionado el primer ítem que se va a suministrar. Responder a la primera pregunta (*¿cuál es el primer ítem que se va a administrar al examinado?*) no es trivial en este caso, porque para saber qué pregunta conviene presentar al inicio de un TAI es necesario disponer de una estimación inicial de la habilidad del evaluando al que se le va a aplicar el test. Es evidente que cuanta más información previa se tenga del examinado, más precisa podrá ser la estimación de partida, y en consecuencia, antes se alcanzará la convergencia con la habilidad real del alumno. Cualquier dato referente al curso, nivel educativo o edad del evaluando, así como la habilidad demostrada en otros tests, puede ser muy útil para obtener una mejor estimación inicial. En contextos reales en los que no se dispone de información previa del alumno se suele realizar una prueba de acceso, permitir que el propio examinado elija su nivel inicial, o seleccionar al azar un ítem de dificultad media.

1.2. ¿Cómo continuar?

En todo algoritmo de aplicación de un TAI debe establecerse un método que resuelva la segunda de las preguntas (*¿qué ítem se va a administrar después de cada respuesta?*). El algoritmo que define el funcionamiento de un TAI es un procedimiento iterativo que, partiendo de la estimación provisional de la habilidad del examinado, evalúa todos los ítems que no han sido utilizados aún en la administración del test, con el fin de seleccionar y administrar el mejor con respecto a esta estimación de habilidad. Tras analizar la respuesta que el alumno ha dado al ítem elegido, se computa una nueva estimación de habilidad (presumiblemente más cercana al valor real, porque para calcularla se ha considerado las respuestas a todos los ítems utilizados hasta el momento, incluido el último) y la precisión asociada. La eficiencia de un TAI depende de estos dos procesos complementarios y estrechamente vinculados, a saber, del método estadístico utilizado para estimar la habilidad y del criterio de selección de ítems.

El **procedimiento de estimación de habilidad** es el encargado de decidir cuál es el valor de la escala de habilidad que más se ajusta al patrón de respuestas observado. Es fundamental que las estimaciones de habilidad que realice el TAI sean precisas, pues en ellas se basa el correcto funcionamiento del test. Destacan dos familias de métodos de estimación: la máximo verosímil condicionada (Lord, 1980), que consiste en maximizar la función de verosimilitud del patrón de respuestas, generalmente mediante algún procedimiento numérico iterativo como el de Newton-Raphson o el algoritmo EM; y la bayesiana (Owen, 1975), que añade a la función de verosimilitud cierta información acerca de la distribución a priori de la habilidad de la población. Para evaluaciones de más de veinte o treinta ítems apenas se encuentran diferencias entre ambos procedimientos (van der Linden y Pashley, 2000).

Por su parte, el **procedimiento de selección del siguiente ítem** a administrar es el proceso responsable de que el TAI sea adaptativo, pues, dependiendo de cuál sea el valor estimado para la habilidad del alumno, escogerá uno u otro ítem para presentar. Entre los criterios para la selección sucesiva de ítems destacan dos por ser los que habitualmente se utilizan para elegir el ítem más adecuado en cada momento: por un lado, los de máxima información (Lord, 1980), consistentes en seleccionar en cada momento, de entre las que aún no se han administrado, la pregunta cuya función de información (fácilmente computable porque sólo depende de los parámetros del ítem) proporciona el mayor valor para el nivel de habilidad estimado hasta el momento, o lo que viene a ser lo mismo, el que minimiza el error estándar, y en consecuencia maximiza la precisión, para la estimación actual de la habilidad del evaluando; y por otro lado, la familia de criterios de selección bayesiana (Owen, 1975), que se basan en actualizar, cada vez que el examinado responde un ítem, además de la estimación actual de la habilidad del sujeto, el valor de la información observada en dicho punto de la escala y la distribución posterior y varianza de habilidades, no sólo para el ítem actual, sino también para todas las respuestas anteriores. De este modo, los procedimientos de selección bayesiana eligen el ítem que minimiza la varianza de la distribución a posteriori de la habilidad. Ambos métodos (máxima información y bayesiano) promueven que aquellos ítems con mayor poder de discriminación sean elegidos una y otra vez para formar parte de los TAI, lo que merma la seguridad del test, y en consecuencia su validez. Actualmente cualquier algoritmo de selección incorpora un mecanismo para controlar la tasa de exposición de los ítems, con el fin de reducirla en el caso de ítems sobreexpuestos e incrementarla en el de los infrautilizados. Se distinguen dos tipos de procedimientos para controlar la exposición de ítems: por una parte, los que seleccionan un ítem al azar entre los más informativos, que se caracterizan por su sencillez de implementación y facilidad de comprensión, y por otra, los que asignan a cada ítem parámetros específicos de control. Algunos TAI también incorporan otra serie de restricciones en la selección de ítems, además de las asociadas a la sobreexposición, tales como las relativas a la homogeneidad o balanceo de contenidos.

1.3. ¿Cómo terminar?

Además del banco de ítems calibrados según un modelo de la TRI, el algoritmo de selección de ítems y el método para estimar la habilidad, es necesario disponer también de un criterio de finalización de la aplicación que determine cuándo ha de terminar el proceso iterativo de administración de ítems, esto es, que dé respuesta a la tercera y última de las preguntas (*¿cuándo se deja de administrar ítems?*). Existen varias alternativas para decidir cuándo dar por finalizada la administración del TAI, entre las que cabe destacar el criterio de longitud fija, para el que el test termina cuando se ha utilizado un número prefijado de

ítems; y el de longitud variable, que dice que el TAI finaliza cuando se ha alcanzado una precisión concreta en la estimación de la habilidad.

Otro tipo de factores, como el tiempo máximo de administración o la detección de patrones aberrantes de respuesta, pueden afectar a la finalización de un TAI. Puede incluso darse el caso de que un algoritmo combine varios criterios de parada. Desde un punto de vista práctico, puede decirse que el criterio de finalización a implementar dependerá del propósito del test, las características del banco de ítems y las restricciones que vengan impuestas por el entorno de aplicación.

A la hora de presentar la evaluación, cabe destacar que la proporción de aciertos, que para la TCT es un buen indicador de la habilidad de los evaluandos, en el caso de los TAI siempre suele rondar el 50%, por lo que no resulta conveniente puntuar las pruebas según este criterio.

5. CONCLUSIONES

Actualmente existen multitud de programas estandarizados de evaluación que incluyen tests en versión adaptativa basados en la TRI, como el National Assessment of Educational Progress (NAEP), el Test of English as a Foreign Language (TOEFL), el Scholastic Assessment Test (SAT), la pionera Armed Service Vocational Aptitude Battery (ASVAB) o el Graduate Record Examination (GRE), y el catálogo de ámbitos en los que a día de hoy se utiliza evaluación adaptativa es muy extenso: medición de aptitudes intelectuales, aptitudes musicales, selección de personal, pruebas de admisión en centros educativos o militares, evaluación curricular, o exámenes de licenciatura o certificación, por citar algunos. En 1990 sólo se administraron unos pocos cientos de TAI, pero esta cifra se incrementó en 1999 hasta superar el millón de aplicaciones, de ahí que el crecimiento en la administración de TAI se estima exponencial(Wainer, 2000).

Los TAI proporcionan innumerables ventajas frente a los procedimientos clásicos de evaluación, pero hay que tener en cuenta que su implantación no resulta fácil, pues el mero hecho de tener que calibrar el banco de ítems a partir de grandes muestras de sujetos, no siempre resulta viable e incluso ha llegado a frustrar muchos proyectos e iniciativas dirigidas hacia la construcción de este tipo de tests. En definitiva, el proceso de desarrollar un TAI es más complejo que redactar un conjunto de ítems en un soporte informático y administrarlos mediante algún tipo de software, pudiendo llegar a requerir varios años de trabajo. Según Wainer (2000) la aplicación de tests adaptativos sólo resulta auténticamente útil bajo determinadas circunstancias, a saber, cuando el rasgo es muy difícil de medir sin un ordenador; cuando el test ha de ofrecerse de forma continuada y no sólo unas pocas veces al año; y cuando se tiene interés en obtener el nivel de habilidad correcto, y no sólo en lograr una puntuación más alta que la que esté establecida. En este sentido, los sistemas de e-learning son más que adecuados para incluir evaluaciones mediante TAI, habida cuenta de la tendencia actual de integrar medición, evaluación, diagnóstico, instrucción y aprendizaje en un mismo contexto. No en vano, es destacable el interés que los desarrolladores de sistemas educativos informatizados parecen mostrar en este paradigma de evaluación. Así, el sistema hipermedia adaptativo para el aprendizaje de idiomas Hezinet-BOGA dispone de un banco de ítems que se está calibrando mediante un novedoso método de aplicar los subtests de anclaje a través de Internet, con el objetivo de realizar los tests de ingreso en el sistema mediante TAI (López-Cuadrado, 2003); el sistema inteligente de evaluación SIETTE (Millán, Trella, Pérez de la Cruz y Conejo, 1999), cuya integración en el curso de cálculo y álgebra LeActiveMath (www.leactivemath.org) se está estudiando, incorpora las redes bayesianas como marco de la evaluación computerizada y también considera la administración de TAI basados en TRI, si bien

su banco de ítems no ha sido calibrado estadísticamente, sino que han sido expertos en la materia quienes han asignado a su juicio los valores de los parámetros de los ítems; el sistema de aprendizaje de matemáticas MATHCAT (Verschoor y Straetmans, 2000), que se utiliza en Holanda para la enseñanza de adultos, y en su versión *-bo* en educación secundaria, cuenta con sendos bancos de 500 y respectivamente 600 ítems calibrados según la TRI; y el Brazilian Portuguese Computerized Adaptative Test (http://www.ku.edu/~brasilis), que se utiliza para evaluar a los estudiantes de portugués brasileño en diversas áreas del conocimiento de la lengua, en combinación con un sistema de aprendizaje por radio que emite ininterrumpidamente, entre otros contenidos, canciones brasileñas e información histórica y cultural.

BIBLIOGRAFÍA

ARMENDARIZ, A. J., LÓPEZ-CUADRADO, J., TAPIAS, A., VILLAMAÑE, M., Sanz-Lumbier, S. Y Sanz-Santamaría, S. (2003). Learning Environments Should Follow Standards: ELSA Does. World Conference On E-Learning In Corporate, Government, Healthcare, & Higher Education (E-Learn 2003), Phoenix, Arizona (USA), Association For The Advancement Of Computing In Education.

BAKER, F. B. (1992). Item Response Theory: Parameter Estimation Techniques. Marcel Dekker. New York (USA).

BIRNBAUM, A. (1968). Some Latent Trait Models And Their Use In Inferring An Examinee's Ability. Statistical Theories Of Mental Test Scores. F. M. Lord Y M. R. Novick (Eds). Addison-Wesley Pp. Chapters 17-20. Reading (USA).

BOCK, R. D. Y AITKIN, M. (1981). Marginal Maximum Likelihood Estimation Of Item Parameters. An Application Of An EM Algorithm. Psychometrika 35.

HAMBLETON, R. K. Y SWAMINATHAN, H. (1985). Item Response Theory : Principles And Applications. Kluwer-Nijhoff Publishing. Boston (USA).

HICKS, M. (1989). The TOEFL Computerized Placement Test: Adaptive Conventional Measurement (TOEFL Research Report No. 31). Princeton, New Jersey (USA), Educational Testing Service.

KOLEN, M. J. Y BRENNAN, R. L. (1995). Test Equating: Methods And Practices. Springer-Verlag. New York (USA).

LÓPEZ-CUADRADO, J., PÉREZ, T. A., ARRUABARRENA, R., Vadillo, J. Á. Y Gutiérrez, J. (2002). Generation Of Computerized Adaptive Tests In An Adaptive Hypermedia System. Educational Technology - Information Society And Education: Monitoring A Revolution. A. Méndez Vilas, J. A. Mesa González Y I. Solo De Zaldívar (Eds). Junta De Extremadura (CECT). 2 Pp. 674-678. Badajoz (España).

LÓPEZ-CUADRADO, J. (2003). The CAT Is Out Of The Bank. Advances In Technology-Based Education: Towards A Knowledge-Based Society. A. Méndez Vilas, J. A. Mesa González Y J. Mesa González (Eds). Junta De Extremadura (CECT). 3 Pp. 1832-1836. Badajoz (España).

LÓPEZ-CUADRADO, J., ARMENDARIZ, A. J. Y PÉREZ, T. A. (2003). ADISTI: An Authoring Tool For Creating And Managing Exercises In E-Learning Systems. Advances In Technology-Based Education: Towards A Knowledge-Based Society. A. Méndez Vilas, J. A. Mesa González Y J. Mesa González (Eds). Junta De Extremadura (CECT). 3 Pp. 1555-1559. Badajoz (España).

LORD, F. M. (1952). A Theory Of Test Scores. Psychometric Monograph 7.

LORD, F. M. (1980). Applications Of Item Response Theory To Practical Testing Problems. Lawrence Erlbaum Associates. Hillsdale, New Jersey (USA).

MILLÁN, E., TRELLA, M., PÉREZ De La CRUZ, J. L. Y CONEJO, R. (1999). Uso De Redes Bayesianas En Test Adaptativos Computerizados. Congreso Nacional De Informática Educativa (CONIED99), Puertollano, Ciudad Real (España).

MISLEVY, R. J. Y BOCK, R. D. (1990). BILOG 3. Scientific Software International. Mooresville (USA).

MUÑIZ, J. (1997). Introducción A La Teoría De Respuesta A Los Ítems. Ediciones Pirámide. Madrid (España).

NOVICK, M. R. (1966). The Axioms And Principal Results Of Classical Test Theory. Journal Of Mathematical Psychology 3.

OLEA, J., PONSODA, V. Y PRIETO, G., Eds. (1999). Tests Informatizados: Fundamentos Y Aplicaciones. Colección "Psicología". Madrid (España), Ediciones Pirámide.

OWEN, R. J. (1975). A Bayesian Sequential Procedure For Quantal Response In The Context Of Adaptive Mental Testing. Journal Of The American Statistical Association 70.

RENOM, J. Y DOVAL, E. (1999). Tests Adaptativos Informatizados: Estructura Y Desarrollo. Tests Informatizados: Fundamentos Y Aplicaciones. J. Olea, V. Ponsoda Y G. Prieto (Eds). Ediciones Pirámide Pp. 127-162. Madrid (España).

SANTISTEBAN, C. Y Alvarado, J. M. (2001). Modelos Psicométricos. UNED Ediciones. Madrid (España).

SEGALL, D. O. Y Moreno, K. E. (1999). Development Of The Computerized Adaptive Testing Version Of The Armed Services Vocational Aptitude Battery. Innovations In Computerized Assessment. F. Drasgow Y J. B. Olson-Buchanan (Eds). Lawrence Erlbaum Associates Pp. 35-65. Mahwah, New Jersey (USA).

VAN DER LINDEN, W. J. Y Pashley, P. J. (2000). Item Selection And Ability Estimation In Adaptive Testing. Computerized Adaptive Testing: Theory And Practice. W. J. Van Der Linden Y C. A. W. Glas (Eds). Kluwer Academic Publishers Pp. 1-25. Dordrecht (The Netherlands).

VERSCHOOR, A. J. Y Straetmans, G. J. J. M. (2000). MATHCAT: A Flexible Testing System In Mathematics Education For Adults. Computerized Adaptive Testing: Theory And Practice. W. J. Van Der Linden Y C. A. W. Glas (Eds). Kluwer Academic Press Pp. 101-116. Dordrecht (The Netherlands).

WAINER, H. Y Mislevy, R. J. (1990). Item Response Theory, Item Calibration And Proficiency Estimation. Computerized Adaptive Testing: A Primer. H. Wainer (Eds). Lawrence Erlbaum Associates Pp. 65-102. Hillsdale, New Jersey (USA).

WAINER, H. (2000). Cats: Whither And Whence. Psicológica 21.

WINGERSKY, M. S. (1983). LOGIST: A Program For Computing Maximum Likelihood Procedures For Logistic Test Models. Applications Of Item Response Theory. R. K. Hambleton (Eds). Educational Research Institute Of British Columbia. Vancouver (Canada).

Contactar

Revista Iberoamericana de Educación

Principal OEI