Revisión de las evaluaciones adaptativas computarizadas (CAT)*

Ruber López**
Paul Sanmartín****
Fernando Méndez*****
Universidad Simón Bolívar, Barranquilla

Recibido: 26 de agosto de 2013 Aceptado: 10 de octubre de 2013

Review of computerized adaptive testing (CAT)

Palabras clave:
Evaluaciones Adaptativas Computarizadas (CAT), Dicotómico, Politómico, TRI, CCI.

Resumen
En este artículo presentamos una revisión de las Evaluaciones Adaptativas Computarizadas. A diferencia de los test convencionales, estas evaluaciones plantean un examen adaptado a las necesidades y capacidades de cada uno de los evaluados, lo cual redunda en una mejor experiencia para el evaluado y en una mayor precisión del resultado. Las evaluaciones adaptativas se fundamentan en la teoría de respuesta a ítems, que define las directrices y condiciones para que este tipo de pruebas sea posible. A partir de esta teoría, surgen distintos modelos que permiten modelar diferentes rasgos de los evaluados y la relación de estos con la probabilidad de acertar un ítem dado. Para llevar a cabo el proceso de evaluación, un test adaptativo debe estar conformado de un banco de ítems, un método que permita la selección de estos y un criterio de terminación. Todos estos componentes articulan la prueba y la ayudan a concretarse adecuadamente.

Key words:
Computerized Adaptive Testing (CAT), Dichotomous, Polytomous, TRI, CCI.

Abstract
This paper is a review of the Computerized Adaptive Testing Process. Unlike conventional tests, these assessments propose a test adapted to every examinee's needs and capabilities, which results in a better experience for those assessed and a more accurate score. Adaptive assessments are based on item response theory, which defines the guidelines and conditions for such tests to be carried out. From this theory, different models that allow the repositioning of different traits of the examinees and their relationship with the probability to succeed in a given item, arise. To complete the assessment process, an adaptive test should consist of a set of items, a method which allows the selection of these and a termination criterion. All the aforementioned components articulate the test and help to properly materialize it.

Referencia de este artículo (APA): López, R., Sanmartín, P. & Méndez, F. (2014). Revisión de las evaluaciones adaptativas computarizadas (CAT). En Revista *Educación y Humanismo, 16*(26), 27-40.

* Artículo vinculado al proyecto Algoritmos para Evaluaciones Adaptativas Computarizadas. Grupo de investigación Innovación Tecnológica y Salud.

** Ingeniero de Sistemas. Estudiante de Maestría en Ingeniería de Sistemas, Universidad Simón Bolívar, Barranquilla. Correo electrónico: rlopez@unisimonbolivar.edu.co
*** Docente investigador, Universidad Simón Bolívar, Barranquilla. Correo electrónico: psanmartin@unisimonbolivar.edu.co
**** Docente investigador, Universidad Simón Bolívar, Barranquilla. Correo electrónico: fmendez1@unisimonbolivar.edu.co

Introducción

Durante mucho tiempo, las evaluaciones se realizaron principalmente en formatos a papel y se han enfocado en la evaluación de desempeño. Pero, desde finales de 1980, con la rápida diseminación de computadores personales en la educación, los formatos de evaluación se han adecuado a las computadoras. Esto último tiene muchas ventajas. Por ejemplo, ofrece la posibilidad de pruebas bajo demanda, esto es, que un evaluado estaría listo para realizar una prueba en cualquier lugar y en cualquier momento. Además, el poder de los computadores modernos y la habilidad de integrar múltiples recursos multimedia puede ser usado para crear formatos de preguntas innovadores y ambientes de prueba más realistas (Van der Linden & Glas, 2010).

Las evaluaciones adaptativas constituyen el siguiente nivel de estas pruebas realizadas en un ambiente computacional, y básicamente muestran un esquema en el que las preguntas no siguen un patrón fijo, sino que, por el contrario, se amoldan a las necesidades de los evaluados a partir de distintos aspectos. Este proceso de evaluación adaptativo hace uso de diversos mecanismos. Su principal fundamento es la teoría de respuesta al ítem, y a partir de esta, se pueden incorporar técnicas como: los métodos bayesianos, lógica difusa, árboles de decisión, entre muchos otros más que hacen posible la generación de este tipo de modelos (Gardner-Medwin, 1995; Petersen *et al.*, 2011; Van der Linden & Glas, 2001; Yen, Ho & Chen, 2006).

Marco general

Teoría Clásica del Test (TCT)

Tiene sus orígenes en los procedimientos promovidos por Galton, Pearson, Spearman, y Thorndike. En esta teoría, el puntaje obtenido en una prueba refleja tanto puntaje verdadero como puntaje de error. Concretamente, el puntaje de la prueba puede ser expresado en la siguiente ecuación:

Puntaje observado = Puntaje verdadero + error

La primera definición que se encuentra en la fórmula es el puntaje observado, el cual hace referencia al puntaje obtenido después de haber realizado la prueba. Por su parte, el puntaje verdadero se define como la cantidad que se obtendría si el evaluado fuese sometido a la prueba un número infinito de veces sin que fuera afectado por situaciones como confusión, fatiga u otro aspecto semejante. El último término en la expresión es el error, definido como la diferencia entre el puntaje verdadero y el puntaje observado. El error no está correlacionado con el puntaje verdadero y el puntaje observado, y se encuentra distribuido normal y uniformemente sobre el puntaje verdadero. Debido a que su influencia es aleatoria, se espera que la cantidad promedio de error a través de varios intentos de la prueba sea cero (De Ayala, 2003; Muñiz, 2010; Weiner, Freedheim, Schinka, Naglieri & Velicer, 2003).

Aunque la teoría clásica del test plantea soluciones a distintos aspectos, presentaba algunas debilidades y fallas en otros. Al respecto, se

puede destacar la invariancia de las mediciones y las propiedades de los instrumentos de medida (Muñiz, 2010). Esto hace referencia a que si se aplica, por ejemplo, una prueba que mide la inteligencia a dos personas, teóricamente, no es posible comparar los resultados de estos individuos, es decir, establecer cuál es más inteligente. Para la solución de este y otros aspectos, surgió la Teoría de Respuesta a Ítem (TRI), mediante la cual se puede hacer lo mismo que con la teoría clásica del test pero de una mejor manera, además de otras cosas adicionales (De Ayala, 2003; Molenaar, 1995).

Teoría de Respuesta a Ítems (TRI)

Constituye un nuevo enfoque en psicometría, y, como se ha dicho, se encamina a contrarrestar algunas limitaciones de la teoría clásica de los test. En este caso, la evaluación que se realiza se basa en aspectos tales como nivel en el tema, la inteligencia de la persona o incluso rasgos de su personalidad (Guzman, Conejo & Perez, 2007; Ho, 2010). Se puede decir que esta teoría se fundamenta en dos principios (Hambleton, Swaminathan & Rogers, 1991; Linacre & Hambleton, 1996):

- El desempeño de un evaluado en un ítem de una prueba puede ser explicado por medio de un conjunto de factores llamados rasgos, rasgos latentes o habilidades.
- La relación entre el desempeño de un evaluado en un ítem y el conjunto de rasgos subyacentes del desempeño de un ítem, puede ser descrita por una función incremental monótona o curva característica del ítem.

La Tabla 1 muestra, precisamente, un análisis comparativo entre la teoría clásica del test y la teoría de respuesta a ítems (Hambleton & Jones, 1993; MacDonald & Sampo, 2002).

Tabla 1. Análisis comparativo TCT y TRI

	Teoría clásica de los test	Teoría de respuesta a ítem
Modelo	Lineal	No lineal
Nivel	Test	Ítem
Suposición	Débil	Fuerte
Relación ítem-habilidad	No especificado	Funciones características de los ítems
Habilidad	El puntaje de test o puntaje real estimado son reportados en la escala del puntaje del test (o transformado a esta escala)	Puntajes de habilidad son reportados en la escala $-\infty$ a $+\infty$ (o una escala transformada)
Invariancia de estadísticas del ítem y del individuo	No - Los parámetros ítem e individuo son muestras dependientes	Sí - Los parámetros ítem e individuo son muestras independientes.
Ítem estadísticos	p, r	b, a y c (para el modelo de tres parámetros), además de las funciones correspondientes de la información de los ítems.
Tamaño de la muestra	200 a 500 (en general)	Depende en el modelo TRI, pero muestras más grandes, generalmente más de 500.

Fuente: Hambleton & Jones, 1993; MacDonald & Sampo, 2002.

La asunción clave en que se basan los modelos de TRI es que existe una relación entre los valores que miden los ítems y la probabilidad de acertar estos, lo cual se denomina función de curva característica del ítem (CCI) (Muñiz, 2010).

Curva característica de un ítem

En situaciones de medición, ya sea en el ámbito psicológico o educativo, se encuentra una intrínseca variable de interés. Esta variable es entendida intuitivamente como inteligencia. En los ámbitos académicos, se suelen utilizar términos descriptivos como habilidad de lectura y habilidad aritmética. La teoría de respuesta a ítem define cada uno de estos términos como rasgo latente (Baker, 2001).

Una suposición razonable de la TRI es que cada evaluado que responde un ítem de una prueba posee una cantidad de habilidad determinada. Por lo cual, se puede considerar que cada evaluado tiene un valor numérico, un puntaje, que lo sitúa en una escala de habilidad. Dicha habilidad se denota con la letra griega theta, Θ. Para cada nivel de habilidad habrá una probabilidad de que un evaluado con cierto nivel de habilidad dé una respuesta correcta en el ítem. Esta probabilidad se denota como P(Θ). En un ítem de una prueba normal, esta probabilidad será pequeña para una habilidad baja, y grande para evaluados con alta habilidad. La Figura 1 muestra un ejemplo claro de una curva característica (Edelen & Reeve, 2007; Hambleton & Swaminathan, 1985).

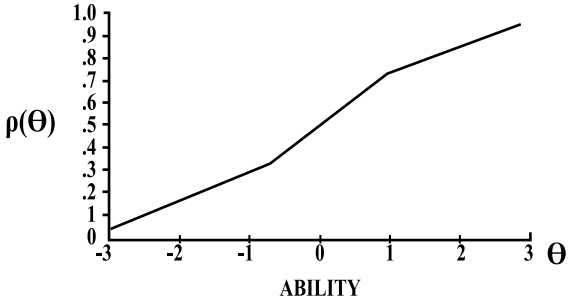

Figura 1. Curva característica del ítem. Representa la probabilidad que un evaluado con cierto nivel de habilidad dé una respuesta correcta en el ítem

Fuente: (Baker, 2001)

Modelos TRI para ítems dicotómicos

Para empezar, se debe tener claro que un ítem dicotómico es aquel cuya respuesta solamente puede ser calificada como correcta o incorrecta. De modo que no hay posibilidad de otro nivel de apreciación en la respuesta. Habiendo aclarado esto y adentrándonos en los modelos TRI para ítems dicotómicos, estos se pueden clasificar según la cantidad de parámetros que se deseen tener en cuenta. Los más utilizados son: dificultad del ítem, representado con la letra *b*; discriminación del ítem, con la letra *a* y el azar o pseudo-azar, por la letra *c*. El parámetro dificultad define qué tan fácil es el ítem, y esto se hace representándolo en una escala que generalmente va de -3 a 3. El parámetro discriminación describe la extensión en la cual un ítem distingue entre evaluados con una habilidad justo por encima o por debajo del nivel de dificultad del ítem actual, teóricamente su valor debería estar entre 0 e infinito, pero por lo general es definido entre 0 y aproximadamente 2. El parámetro de pseudo-azar define la posibilidad que tiene un evaluado de responder correctamente a causa de suerte. Su valor se define en la misma escala que la probabilidad (Hambleton & Swaminathan, 1985).

Habiendo visto los distintos parámetros considerados en los modelos TRI para ítems dicotómicos, resulta pertinente identificar cómo se configura cada uno de estos. El más parametrizado es el modelo logístico de tres parámetros, que contempla las tres variables mencionadas

anteriormente. De ahí en adelante, los modelos logísticos de dos y un parámetros son simplificaciones de este (suprimen algunos parámetros). A continuación veremos cada uno de estos más a fondo.

a) Modelo logístico de tres parámetros

Tiene en cuenta los tres parámetros: dificultad (b), discriminación (a) y pseudo-azar (c) (Birnbaum, 1968). En este modelo, se entiende la probabilidad de una respuesta correcta para una persona j, en un ítem i dado el nivel de habilidad θ_j.

b) Modelo logístico de dos parámetros

Tiene en cuenta únicamente dos parámetros: dificultad (b) y discriminación (a) (Birnbaum, 1968). En este modelo, también se entiende la probabilidad de una respuesta correcta para una persona j en un ítem i dado el nivel de habilidad θ_j.

c) Modelo logístico de un parámetro

También conocido como modelo Rasch (Molenaar, 1995), solo tiene en cuenta un parámetro: dificultad (b) (Birnbaum, 1968). En este modelo, se entiende la probabilidad de una respuesta correcta (x_i=1) para una persona j en un ítem i, dado el nivel de habilidad θ_j.

Modelos TRI para ítems politómicos

Un ítem politómico, a diferencia de uno dicotómico, es aquel cuya respuesta contempla más de dos niveles de validez. Estas respuestas pueden ser evaluadas más allá de un simple correcto o incorrecto, por ejemplo, se puede tener: malo, medio y excelente. Ahondando en los modelos TRI para ítems politómicos, se encuentra que la mayoría de estos son extensiones de los modelos TRI dicotómicos. Estos modelos, más que usar un simple parámetro de dificultad, hacen uso de categorías de dificultad para describir la posibilidad de responder en cada una de estas.

Thissen y Steinberb (1986) recopilaron varios modelos TRI politómicos, y los clasificaron en dos tipos: modelos de diferencia y modelos de división por el total. En los modelos de diferencia, como son el modelo de respuesta graduada (Samejima, 1969) y el planteado por Muraki (rating scale model), se requiere un proceso de dos estados para determinar la probabilidad de recibir una categoría de puntaje. Por su parte, los modelos de división por el total, como el modelo de crédito parcial (Masters, 1982) y el modelo de crédito parcial propuesto por Muraki (1992), la probabilidad de responder en una categoría se encuentra determinado por un cálculo directo. Entre estos, los más comúnmente usados son: el modelo de respuesta graduada, el modelo de crédito parcial, y el modelo de crédito parcial generalizado, los cuales veremos a continuación.

a) Modelo de respuesta graduada –Graded Response Model (GRM)

Este modelo fue propuesto por Samejima (1969), y es una adaptación del modelo logístico de dos parámetros, pero enfocado en ítems politómicos. En este modelo, a cada categoría de respuesta se le da, en primera instancia, un carácter dicotómico, para calcular la probabilidad de que un evaluado j reciba un puntaje de

categoría "x" o superior en el ítem i, dada una habilidad θ_j (Samejima, 2008; Ho, 2010; Ostini & Nering, 2005).

b) Modelo de crédito parcial –Partial Credit Model (PCM)

Este modelo fue propuesto por Masters (1982). Plantea que los ítems no se diferencian en cuanto a su discriminación y que el azar es irrelevante. Este modelo parte de la base del modelo Rasch, pero adaptado para ítems politómicos. Se puede decir que es una adaptación del modelo logístico de dos parámetros pero enfocado en ítems politómicos. En este modelo se plantea la probabilidad que un evaluado j reciba un puntaje de categoría "x" en el ítem i, dado una habilidad θ_j (Fox, 2008; Ho, 2010; Ostini & Nering, 2005).

c) Modelo de crédito parcial generalizado –Generalized Partial Credit Model (GPCM)

Este modelo fue propuesto por Muraki (1992). Su base es el modelo logístico de dos parámetros y plantea la implementación en ítems politómicos. En él, se plantea la probabilidad de que un evaluado j reciba un puntaje de categoría "x" en el ítem i, dada una habilidad θ_j (Hayes, 2012; Ho, 2010; Ostini & Nering, 2005).

Teoría de Respuesta a Testlet (TRT)

Básicamente, un testlet es "un conjunto de ítems relacionados con un solo estímulo que es desarrollado como una unidad y contiene un número fijo de rutas que el evaluado puede seguir" (Wainer & Kiely). Tenemos como ejemplo los fragmentos de lecturas, de los cuales se desprenden un conjunto de preguntas. De igual manera, encontramos los problemas matemáticos sobre los que se plantean diferentes ítems referidos a la solución.

Uno de los principales retos de este tipo de pruebas es el de la Dependencia Local de los Ítems (DLI), principio básico de la teoría de respuesta a ítems, y esto es debido a que no se puede ver cada ítem por separado, ya que de alguna manera se relacionan con respecto al rasgo que desean evaluar.

Como respuesta al problema antes mencionado, Wainer and Lewis (1990) han comentado tres posibles soluciones. La primera consiste en revisar el formato de la prueba, de tal manera que solo haya un ítem conectado con un estímulo común. Pero esta aproximación acarrea otro problema, y es el uso ineficiente del tiempo de la prueba y esfuerzo del evaluado. Por otro lado, está la opción de ignorar el problema de DLI y calibrar los datos de respuesta de los ítems usando los conocidos modelos TRI dicotómicos. El inconveniente, en este caso, es que la ignorancia de la DLI lleva a una sobreestimación de la información de la prueba. Una tercera aproximación es la de modelar el testlet como si fuera un simple ítem politómico. Esta ha demostrado ser una solución efectiva para el problema de la DLI (Thissen, Steinberg & Mooner, 1989; Yen; 1993; Wainer, 1995). Sin embargo, alternativamente, el problema de la DLI puede ser manejado por medio de los modelos de medición planteados por la teoría de respuesta a testlet (TRT) (Wainer, Bradlow & Wang, 2007). La TRT no solo re-

suelve el problema de la DLI, sino que también mantiene los ítems como unidades de medición (Ho, 2010).

1. The 2PL testlet response theory model. El modelo logístico de dos parámetros de la teoría de respuesta a testlet (2PL-TRT) es el modelo inicial desarrollado por Bradlow *et al.* (1999). Funciona como base de modelos TRT más complejos y es una modificación del modelo logístico de dos parámetros (Birnbaum, 1968) para justificar la dependencia local entre los ítems dentro del mismo testlet. Para el 2PL-TRT, la probabilidad de que una persona *i* con una habilidad θ_i responda correctamente un ítem j dentro del testlet, *d(j)*, es designado como $P_{ij}(y=1 \mid \theta_i)$, y se expresa de la siguiente manera:

$$P_{ij}\,(y_j = 1|\theta_i) = \frac{exp\,(a_j(\theta_i - b_j - Y_{id\,(j)}))}{1 + exp\,(a_j(\theta_i - b_j - Y_{id\,(j)}))}$$

Pruebas adaptativas computarizadas –Computerized Adaptive Testing (CAT)

El objetivo final de una evaluación adaptativa consiste en estimar de manera cuantitativa el nivel de conocimiento del evaluado por medio de un valor numérico. Para este fin, los ítems se colocan secuencialmente, es decir, uno a la vez (Huo, 2009). La presentación de cada ítem y la decisión de finalizar el test son dinámicamente adoptadas, según la respuesta del evaluado. En general, una evaluación adaptativa aplica un algoritmo iterativo, el cual inicia con una estimación inicial del conocimiento del evaluado y cumple los siguientes pasos (Guzman *et al.*, 2007; Sébille *et al.*, 2010; Ware *et al.*, 2003):

- En la lista de ítems disponibles, todos son examinados para determinar cuál funciona mejor en una secuencia que se desarrolla según la estimación del nivel de conocimiento identificado en el evaluado.
- La pregunta es formulada y el evaluado la responde.
- De acuerdo con la respuesta, se realiza una nueva estimación de su nivel de conocimiento.
- El paso inicial y el tercero se repiten hasta que se encuentra un criterio de finalización.

Las CAT se podrían definir como una forma de evaluación computarizada que se basa en su adaptación, dependiendo de distintos factores encontrados en el usuario de la prueba. Hacen uso de diferentes mecanismos para identificar los casos en que la prueba debe adaptarse a las necesidades del usuario y así proporcionarle una que se ajuste a su nivel y aptitudes (Wainer, Dorans, Flauger, Green, Mislevy, 2003).

El proceso de adaptación en las pruebas adaptativas se pueden llevar a cabo por diferentes medios:

- Teoría de respuesta a ítem
- Identificación de modelos de estudiantes
- Información acumulada
- Identificación de niveles y dimensiones cognitivas
- Diferentes algoritmos

Para la clasificación de los ambientes de evaluaciones adaptativas, se toma como referencia a Weiss y Kingsbury (2005), según los cuales los

diferentes esquemas de evaluaciones adaptativas pueden ser comparados, a partir de seis criterios principales (Chajewski, 2011; Weiss & Kingsbury, 2005):

- Modelo de respuesta a ítem
- Un banco (pool) de ítems
- Reconocimiento previo antes de la prueba
- Una regla para la selección de ítems
- Un método de puntuación
- Un criterio de finalización

Modelo de respuesta a ítem

Describe la teoría de evaluación educativa subyacente en un ambiente de prueba adaptativo. Generalmente, estos ambientes adaptativos están basados en la llamada Teoría de Respuesta a Ítem (TRI) (Collins, 1996).

Banco de ítems

Es el conjunto acumulado de ítems que se usan para determinar el nivel de conocimiento con respecto a un tema específico. En primera instancia, un ítem puede ser indexado de acuerdo con el objetivo de aprendizaje que contribuye a evaluar. Este índex es usado para indicar el grado en que es capaz de distinguir la pertenencia o no del conocimiento. Cabe destacar la utilidad de estos índices en el proceso de selección de los ítems, ya que las preguntas demasiado difíciles podrían frustrar al evaluado, mientras que si, por el contrario, son demasiado fáciles, podrían hacer que el evaluado se canse en la prueba (Collins, 1996; Lopez-Cuadrado, Perez, Vadillo, Gutiérrez, 2010; Raykova, Kostadinova, Totkov, 1999).

Reconocimiento previo antes de la prueba

Un conocimiento previo con respecto al nivel de dominio del tema del evaluado podría estar disponible como entrada para la prueba. En este sentido, el instructor o el algoritmo definido para la prueba podrían poseer esta información del evaluado, y estos datos pueden ser usados como conocimiento extra.

Típicamente, las fuentes del conocimiento extra pueden provenir de algún tipo de prueba preliminar, de intentos de evaluación anteriores o de alguna observación que se haga en el evaluado. Inclusive, resultados de pruebas fallidas pueden ser útiles para disminuir el tiempo empleado en identificar el nivel del evaluado.

En caso de que un estudiante presente fortalezas en un tema y falencias en otra área, la prueba puede utilizar esta información de entrada y así enfocarse en la que más necesita el evaluado en esos momentos (Collins, 1996).

Reglas para la selección de los ítems

La selección de los ítems es muy importante para lograr que la dinámica de una prueba adaptativa sea exitosa. Si se quiere determinar rápidamente el nivel de dominio del tema que tiene el evaluado, deben tomarse en consideración distintos factores. Entre los más comunes, se encuentran la discriminación de los índices y el nivel de dificultad. La selección de los ítems debería incluso tomar en consideración el conocimiento actual del evaluado. Otro aspecto importante para evitar el carácter determinístico de

la prueba consiste en agregar algo de aleatoriedad en la rutina de selección (Linacre, 2000; Van der Linden, Glas, 2010).

Método de puntuación

Después que cada ítem es presentado y se provee una respuesta para este por parte del evaluado, se obtiene información adicional y se incorpora en el conocimiento general sobre el evaluado en el sistema. A esto se le llama puntuación.

Generalmente, los ítems de las pruebas tienen respuestas objetivas y es fácil determinar si son correctas o no. A la respuesta obtenida se le aplica el método de puntuación para actualizar el nivel de dominio del tema estimado para el evaluado (Collins, 1996).

Criterio de finalización

Usando los resultados del método de puntuación, el criterio de terminación puede ser definido por la prueba. Cuando un estudiante sobrepasa el puntaje o el criterio definido, la rutina de la prueba puede concluir que el evaluado tiene dominio, ya sea de un tema en particular o de un conjunto de temas (He, 2010). De igual manera, puntajes inferiores al criterio definido pueden inducir a que la prueba determine que el evaluado no posee un dominio adecuado del tema. La seguridad de la terminación de la prueba se basa primeramente en una función del método de puntuación, ya sea estadística o empíricamente (Van der Linden & Glas, 2010).

Trabajos acerca de evaluaciones adaptativas

Muchos trabajos se han realizado en el marco de las evaluaciones adaptativas, la mayoría de ellos utilizando la teoría de respuesta al ítem o métodos bayesianos. A continuación, presentamos algunos ejemplos:

Estimación de habilidad en un sistema de evaluación adaptativa con lógica difusa

Está investigación se centra en el planteamiento de un nuevo modelo llamado Fuzzy Item Response Model (FIRM), que combina la teoría de respuesta al ítem con la teoría difusa en la estimación de habilidades o competencias (Balas-Timar & Balas, 2009).

Sistema de evaluación adaptativa basada en árbol de decisión

Se basa en el método de Bayes. En vez de la teoría de test, propone un nuevo empleo de las pruebas adaptativas según árboles de decisión (Ueno & Songmuang, 2010).

Investigación en sistemas de pruebas adaptativas basados en la teoría de respuesta

Se basa en la teoría de respuesta a ítem pero propone un diseño integrado en el cual se combinan la construcción de un banco de preguntas y el gsm (grading scoring model) (Liu, Ping, Zhiliang, Pan, 2010).

Prueba adaptativa politómica que premia el conocimiento parcial

Propone un sistema de evaluación adaptativa incorporando un esquema de medición de

confianza dentro de un modelo de respuesta graduado (GRM). El resultado muestra qué método politómico es capaz de evaluar el conocimiento parcial de los participantes con menos preguntas y con una validez predictiva más alta (Yen *et al.*, 2006).

Adaptative Testing for Hierarchical Student Models

Presenta una aproximación para modelado de estudiantes en el cual el conocimiento es representado por medio de distribuciones de probabilidad asociadas a árboles de conceptos. Usa un modelo politómico, ya que en este es posible dar un crédito parcial a las respuestas, lo cual brinda mayor información y un diagnóstico más eficiente. Esto enmarcado dentro de la teoría de respuesta a ítem (Guzman *et al.*, 2007).

Utilizing Response Time Distributions for Item Selection in CAT

En este trabajo, los autores proponen dos criterios para la selección de los ítems que utilizan un modelo log-normal para el tiempo de respuesta. El primero modifica el criterio de máxima información para maximizarla por unidad de tiempo. El segundo es una versión inversa de la estratificación del ponderado del tiempo que toma ventaja del modelo de tiempo de respuestas (Fan, Wang, Chang & Douglas, 2012).

Computerized adaptive Test Item Response Times for Correct and Incorrect Pretest and Operational Items: Testing Fairness and Test-Taking Strategies

Este estudio examinó la cantidad de tiempo utilizado en las evaluaciones previas (pretest) e ítems operativos, que son respondidos correcta e incorrectamente por evaluados con diferentes niveles de habilidad cuando toman una evaluación adaptativa computarizada. Los resultados indicaron que evaluados con mayor habilidad gastan más tiempo que los que poseen una habilidad menor (Chang, 2007).

Conclusiones

En este artículo, se presentó una revisión sobre las evaluaciones adaptativas computarizadas. Estas se fundamentan en la *teoría de respuesta a ítems,* que define las directrices y condiciones para que este tipo de pruebas sea posible. A partir de esta teoría, surgen distintos modelos que permiten modelar diferentes rasgos de los evaluados y la relación de estos con la probabilidad de acertar un ítem dado. Pero también encontramos la *teoría de respuesta a testlet*, que es "un conjunto de ítems relacionados con un solo estímulo que es desarrollado como una unidad y contiene un número fijo de rutas que el evaluado puede seguir" (Wainer & Kiely, 1987). Con el potenciamiento de la computación en los últimos años, la teoría de respuesta a ítem y los modelos de evaluación han encontrado nuevas formas de aplicabilidad. Esto ha ayudado a fortalecer y mejorar muchos componentes que intervienen en el proceso de evaluación, ya sea en la forma como se presentan los ítems a los usuarios, los criterios utilizados para su selección, el tiempo de duración de la prueba, entre otros. En este ámbito se han desarrollado las evaluaciones adaptativas, que, como se ha visto, plantean un modelo flexible de evaluación que mejora algunos aspectos

en los que fallaban antiguos modelos, potenciándolos y llevándolos a un nivel más avanzado. Este tipo de evaluaciones ha abierto un mundo de opciones, que en colaboración con diferentes técnicas, como la lógica difusa, las redes neuronales, árboles de decisión y muchas otras más, planteen futuros retos, siempre buscando una mayor fiabilidad y precisión en los resultados arrojados por estas pruebas.

Referencias

Baker, F. B. (2001). The item characteristic curve. En: F. B. Baker, *The basics of item response theory.* Second edition (p. 15). Washington: Heinemann.

Balas-Timar, D. V. & Balas, V. E. (2009). *Ability Estimation in CAT with Fuzzy Logic.* Recuperado el 10 de octubre de 2011 de http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=5342278&contentType=Conference+Publications

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

Bradlow, E., Wainer, H. & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64, 153-168.

Chajewski, M. (2011). MLE vs. bayesian item exposure in non-cognitive type adaptive assessments with restricted item pools: Trait estimation, item selection and reliability. Fordham University. *ProQuest Dissertations and Theses,* 406. Retrieved from http://search.proquest.com/docview/924472386?accountid=45648. (924472386).

Chang, S. (2007). Computerized adaptive test item response times for correct and incorrect pretest and operational items: Testing fairness and test-taking strategies. The University of Nebraska-Lincoln. *ProQuest Dissertations and Theses*, 141-141.

Collins, J. A. (1996). *Adaptive Testing with Granularity.* Recuperado el 15 de abril de 2012 de http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.56.8113

De Ayala, R. J. (2003). *The theory and practice of item response theory.* Portland: Jhon Guildford Press.

Edelen, M. O. & Reeve, B. B. (2007). Applying Item Response Theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*, 5-18. doi: http://dx.doi.org/10.1007/s11136-007-9198-0

Fan, Z., Wang, C., Chang, H.-H. & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics, 37*(5), 655.

Fox, C. (2008). An introduction to the partial credit model for developing nursing assessments. *Journal of Nursing Education, 38*(8), 340-346. Retrieved from http://search.proquest.com/docview/203940111?accountid=45648

Gardner-Medwin, A. R. (1995). *Confidence assessment in the teaching of basic science.* Recuperado el 20 de mayo de

2012 de http://www.researchinlearningtechnology.net/index.php/rlt/article/view/9597/11205

Guzman, E., Conejo, R. & Perez, J. L. (2007). *Adaptive testing for hierarchical student models*. Recuperado el 20 de octubre de 2011 de http://dx.doi.org/10.1007/s11257-006-9018-1

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Hambleton, R. K. & Jones, R. W. (1993). *Comparison of classical test theory and item response theory and their applications to test development* (p. 258). Educational measurement: issues.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). Concepts, Models, and Features. En R. K. Hambleton, H. Swaminathan, H. J. Rogers, *Fundamentals of Item Response Theory* (p. 7). California: Sage Publications, Inc.

Hayes, H. (2012). A generalized partial credit FACETS model for investigating order effects in self-report personality data. Georgia Institute of Technology. *ProQuest Dissertations and Theses*, 184. Retrieved from http://search.proquest.com/docview/1114903184?accountid=45648. (1114903184)

He, W. (2010). Optimal item pool design for a highly constrained computerized adaptive test. Michigan State University. *ProQuest Dissertations and Theses*, 157. Retrieved from http://search.proquest.com/docview/815435292?accountid=45648. (815435292)

Ho, T. (2010). A comparison of item selection procedures using different ability estimation methods in computerized adaptive testing based on the generalized partial credit model. The University of Texas at Austin. *ProQuest Dissertations and Theses,* 198. Recuperado de http://search.proquest.com/docview/760034367?accountid=45648. (760034367)

Huo, Y. (2009). Variable-length computerized adaptive testing: Adaptation of the a-stratified strategy in item selection with content balancing. University of Illinois at Urbana-Champaign. *ProQuest Dissertations and Theses,* 82-n/a. Retrieved from http://search.proquest.com/docview/304898277?accountid=45648. (304898277)

Kinsey, T. L. (2003). A comparison of IRT and rasch procedures in a mixed-item format test. University of North Texas. *ProQuest Dissertations and Theses,* 122-122 p.

Zaina, L., Cardieri, M. & Bressan, G. (2011). Adaptive learning in the educational e-lors system: an approach based on preference categories. *IJLT, 6*(4), 341-361.

Linacre, M. J. (2000). *Computer-Adaptive Testing: A Methodology Whose Time Has Come*. Recuperado el 20 de mayo de 2012 de http://rasch.org/memo69.pdf

Linacre, M. J. & Hambleton, R. K. (1996). *Handbook of Modern Item Response Theory.* New York: Springer.

Liu, C., Ping, M., Zhiliang, Z. & Pan, Y. (2010). *Research on computerized adaptive testing system based on IRT.* Recuperado el 15 de octubre de 2011 de http://dx.doi.org/10.1109/FSKD.2010.5569425

Lopez-Cuadrado, J., Perez, T. A., Vadillo, J. A. & Gutiérrez, J. (2010). *Calibration of an item bank for the assessment of Basque language knowledge.* Recuperado el 15 de octubre de 2011 de http://dx.doi.org/10.1016/j.compedu.2010.04.015

MacDonald, P. & Sampo, V. P. (2002). A montecarlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62*(6), 921-943.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

Molenaar, I. W. (1995). Some background for item response theory and the rash model. En G. H. Fischer & I. W., Molenaar, *Rasch models: foundations, recent developments, and applications* (pp. 3-5). New York: Springer-Verlag.

Muñiz, J. (2010, enero). Las teorías de los tests: Teoría clásica y teoría de respuesta a los ítems. *Papeles del psicólogo, 31*. Recuperado el 20 de mayo de 2012 de http://www.papelesdelpsicologo.es

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14*(1), 59-71.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176. doi: http://dx.doi.org/10.1177/014662169201600206

Ostini, R. & Nering, M. (2005). *Polytomous Item Response Theory Models.* California: Sage Publications.

Petersen, M. A., Groenvold, M., Aaronson, N. K., Chie, W., Conroy, T., Costantini, A.,... Young, T. (2011). Development of computerized adaptive testing (CAT) for the EORTC QLQ-C30 physical functioning dimension. *Quality of Life Research, 20*(4), 479-90. doi: http://dx.doi.org/10.1007/s11136-010-9770-x

Raykova, M., Kostadinova, H. & Totkov, G. (1999). *Adaptive test system based on revised Bloom's taxonomy*. Recuperado el 16 de octubre de 2011 de http://doi.acm.org/10.1145/2023607.2023692

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded responses. *Psychometrika Monograph*, 17.

Samejima, F. (2008). Graded response model based on the logistic positive exponent family of models for dichotomous responses. *Psychometrika, 73*(4), 561-578. doi: http://dx.doi.org/10.1007/s11336-008-9071-2

Sébille, V., Hardouin, J., Le Néel, T., Kubis, G., Boyer, F., Guillemin, F. & Falissard, B. (2010). Methodological issues regarding power of classical test theory (CTT) and item response theory (IRT)-based approa-

ches for the comparison of patient-reported outcomes in two groups of patients - a simulation study. *BMC Medical Research Methodology, 10*(1), 24-24. doi: http://dx.doi.org/10.1186/1471-2288-10-24

Thissen, D. & Steinberg, L. (1986). Taxonomy of item response models. *Psychometrika*, *51*(4), 567-577.

Thissen, D., Steinberg, L. & Mooney, J. A. (1989, septiembre). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational.*

Ueno, M. & Songmuang, P. (2010). *Computerized Adaptive Testing Based on Decision Tree*. Recuperado el 10 de octubre de 2011 de http://doi.ieeecomputersociety.org/10.1109/ICALT.2010.58

Van der Linden, W. J. & Glas, C. A. (2001). *Computerized adaptive testing: Theory and practice.* Norwell: Kluwer Academic Publishers.

Van der Linden, W. J. & Glas, C. A. (2010). *Elements of Adaptive Testing.* New York: Springer.

Wainer, H. (1995). Recision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8*(2), 157-187.

Wainer, H. & Kiely, G. (1987). Item Clusters and Computerized Adaptive Testing: A Case for Testlets. *Journal of Educational Measurement, 12*, 241-249.

Wainer, H. & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*(1), 1-14.

Wainer, H., Bradlow, E. T. & Wang, X. (2007). *Testlet response theory and its applications.* New York: Cambridge University Press.

Wainer, H., Dorans, N. J., Flauger, R., Green, B. F. & Mislevy, R. J. (2003). *Computerized adaptive testing: a primer.* Routledge.

Ware Jr, J. E., Kosinski, M., Bjorner, J. B., Bayliss, M. S., Batenhorst, A., Dahlöf, C. G. & Dowson, A. (2003). Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research, 12*(8), 935-52. doi: http://dx.doi.org/10.1023/A:1026115230284

Weiner, I. B., Freedheim, J. R., Schinka, J. Naglieri, J. A. & Velicer, W. F. (2003). *Handbook of Psychology.* New York: Wiley.

Weiss, D. J. & Kingsbury, G. G. (2005). *Application of Computarized adaptive testing to educational problems.* Recuperado el 18 de octubre de 2011 en: http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.1984.tb01040.x/abstract

Yen, C., Yan, Y. & Chong, M. (2006). Correlates of methamphetamine use for Taiwanese adolescents. *Psychiatry and Clinical Neurosciences, 60*(2), 160-167.

Yen, W. M. (1993). Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213.

Yen, H. & Chen (2006). "Taiwan's Economic Development: The Role of Entrepreneurship and Its Incubating Factors", *Global Business & Economics Review-Anthology*, 525-534.