Revista Facultad de Ciencias Universidad Nacional de Colombia, Sede Medellín
V 4 N°2 julio-diciembre de 2015 • ISSN 0121-747X / ISSN-e 2357-5749 • Artículo Investigación • Páginas 58 a 73
DOI: https://doi.org/10.15446/rev.fac.cienc.v4n2.53174

DISEÑO Y VALIDACIÓN DE UN MODELO DE EVALUACIÓN DINÁMICA BASADO EN LA TEORÍA DE RESPUESTA AL ÍTEM[a]

DESIGN AND VALIDATION OF A DYNAMIC ASSESSMENT MODEL BASED ON THE ITEM RESPONSE THEORY

JULIÁN MORENO CADAVID[b], LUIS FERNANDO MONTOYA GÓMEZ[b]

Recibido 22-09-2015, aceptado 11-11-2015, versión final 01-12-2015.
Artículo Investigación

RESUMEN: El objetivo de este artículo es describir un modelo de evaluación basado en un proceso iterativo de selección de ítems a partir de una estimación dinámica del conocimiento del estudiante. La idea tras dicho modelo es lograr un equilibrio entre el progreso demostrado por el estudiante y la dificultad de los ítems que son usados durante el proceso de evaluación. De esta manera se logra evitar de cierta manera dos fenómenos comunes: el tedio en los estudiantes 'avanzados', y la frustración en los 'rezagados'. Para el diseño de dicho modelo se siguió el enfoque de la Evaluación Adaptativa Computarizada usando la Teoría de Respuesta al Ítem y en particular el modelo logístico de tres parámetros. Para validar la propuesta se llevó a cabo un estudio con una muestra de 59 estudiantes en un ambiente educativo real, específicamente dentro de un curso universitario de estructuras de datos. Dicho estudio demostró, por medio de pruebas de hipótesis, no solo un incremento significativo en la precisión de la medición respecto al nivel de conocimiento de los examinados, sino también una disminución en el tiempo promedio de evaluación.

PALABRAS CLAVE: Teoría de respuesta al ítem, evaluación, estimación dinámica, 3PL.

ABSTRACT: The goal of this paper is to describe an assessment model based on an item selection iterative process using a dynamic estimation of the student knowledge. The idea behind such a model is to achieve equilibrium between the progress demonstrated by the student and the items that are used during the assessment process. In this way two common phenomena are avoided: the boredom of 'advanced' students, and the frustration of the 'straggler' ones. For the design of such a model the Computerized Adaptive Testing approach was followed using the Item Response Theory and in particular the logistic model with three parameters. To validate the proposal a study was performed with 59 students in a real educational environment, specifically within a college course in data structures. Such a study demonstrated, through hypothesis tests, not only a significant raise in the measuring accuracy with regard to the knowledge level of the examinees,

[a]Moreno, J. & Montoya, L. (2015). Diseño y validación de un modelo de evaluación dinámica basado en la teoría de respuesta al ítem. *Revista de la Facultad de Ciencias*, 4 (**2**), 58–73.DOI: https://doi.org/10.15446/rev.fac.cienc.v4n2.53174

[b]Departamento de Ciencias de la Computación y la Decisión, Universidad Nacional de Colombia, Medellín.
jmoreno1@unal.edu.co

but also a dropout in the mean assessment time.

KEYWORDS: Item response theory, assessment, dynamic estimation, 3PL.

1. INTRODUCCIÓN

La evaluación es un elemento clave de cualquier proceso educativo, no solo porque permite verificar que los estudiantes alcancen los objetivos curriculares propuestos, sino también porque sirve como herramienta para validar las estrategias pedagógicas empleadas.

En palabras del Ministerio de Educación Nacional de Colombia (2003): " 'Lo que se mide mejora', es el espíritu que anima el proceso de evaluación, que va de la mano con los estándares y con los planes de mejoramiento; entre sí se alimentan, al medir la pertinencia de los conocimientos adquiridos y formular acciones con base en los resultados".
Cuando es llevada a cabo en el contexto de la educación presencial, la evaluación generalmente hace uso de uno o varios mecanismos que permiten una interacción relativamente directa entre el docente y el estudiante: exámenes escritos u orales, presentaciones, discusiones, etc. (Trevitt *et al.*, 2012). En contraste a esta situación, dentro de la educación soportada por Tecnologías de la Información y las Telecomunicaciones (TIC), donde la automatización suele ser uno de los principios guías del diseño instruccional, la evaluación suele ser realizada de manera indirecta mediante pruebas computarizadas, también denominadas informatizadas. Tales pruebas en la mayoría de los casos se componen de un número predefinido de preguntas, seleccionadas de manera determinística o aleatoria, que luego son presentadas a los estudiantes. En ambos casos, educación presencial y a distancia, ocurre un fenómeno y es que, sea por falta de tiempo del docente en el primer caso, o por la naturaleza misma de las pruebas en el segundo, el proceso de evaluación dista mucho de ser personalizado. En otras palabras, todos los estudiantes son tratados de la misma manera, salvo quizá por el componente aleatorio opcional en el segundo caso, sin considerar sus diferencias individuales.

Como una alternativa a estos panoramas la Evaluación Adaptativa Computarizada, o EAC, se diferencia en que su proceso de construcción es dinámico según el progreso del estudiante y en que la cantidad de preguntas no es fija (Weiss & Kingsbury, 1984; Huang, 1996; Eggen & Straetmans, 2000; Guzmán *et al.*, 2007; Van der Linden & Glas, 2010; Thompson & Weiss, 2011; López *et al.*, 2014). En este punto es importante hacer una diferencia entre dos términos que comúnmente son confundidos entre sí: adaptable y adaptativo. Los sistemas que permiten que los usuarios alteren determinados parámetros de configuración para luego adaptarse a dichas alteraciones de una manera explícita son denominados adaptables. Por otro lado, los sistemas que se adaptan a los usuarios de forma implícita a partir de las inferencias realizadas automáticamente sobre las

necesidades de dichos usuarios, se denominan adaptativos (Oppermann & Kinshuk, 1997).

La idea fundamental de la EAC es la aplicación, no de todas las disponibles, sino únicamente de aquellas preguntas que son útiles para estimar el nivel de competencia del estudiante. Como consecuencia de esto, la EAC suele ser más eficiente que los test convencionales en el sentido que provee mediciones más precisas para pruebas de la misma longitud o pruebas más cortas para las mismas medidas de precisión (Ponsoda, 2000; Triantafillou *et al.*, 2008; Özyurt *et al.*, 2012; Chang, 2014).
Y las ventajas no son solo desde el punto de vista del evaluador. De hecho, para el evaluado esta aproximación permite tener una percepción de que la dificultad de la prueba generada se ajusta a su nivel de competencia. Así por ejemplo, si un estudiante responde de manera correcta una pregunta de nivel intermedio, lo más probable es que la siguiente pregunta que se le presente sea de una dificultad mayor. Por el contrario, si ese estudiante responde esa pregunta de forma incorrecta, lo más probable es que la siguiente disminuya en dificultad. Esto no significa en ningún momento que la intención de la EAC sea facilitar por un lado la evaluación a los estudiantes con un desempeño bajo, o por otro dificultársela a aquellos que demuestren dominio sobre el tema evaluado. Lo que la EAC en realidad pretende es evitar la frustración que pueden llegar a sentir los estudiantes que se "bloquean" cuando se ven enfrentados a una evaluación que les resulta difícil, así como el tedio de aquellos que sienten que se les repiten temas que ya han demostrado dominar.

2. DESCRIPCIÓN DEL MODELO

En el nivel de abstracción más alto, el modelo propuesto sigue la estructura general de la EAC, la cual consiste de un algoritmo iterativo con los siguientes pasos (Thissen & Mislevy, 2000):

1. Se selecciona de un banco de ítems de evaluación, es decir, preguntas, el más adecuado para el estudiante según su actual nivel de conocimiento estimado.

2. Se presenta dicho ítem al estudiante, quien lo responde de manera correcta o incorrecta.

3. Se actualiza el nivel de conocimiento estimado de dicho estudiante a partir, no solo de esa respuesta, sino del acumulado de todas las que se haya presentado previamente.

4. Los pasos 1 a 3 se repiten hasta que se alcance un determinado criterio de parada.

Según este algoritmo, los cuatro elementos fundamentales de la EAC son: a) un repositorio de ítems de evaluación, b) un criterio de selección de dichos ítems, c) un procedimiento para estimar el nivel de conocimiento del evaluado, y d) un criterio de terminación de la evaluación. Con respecto al primer elemento, el repositorio de ítems, la consideración más importante a tener en cuenta es que a mayor cantidad de ítems y que estos estén bien descritos, mayor será la capacidad para evaluar

a los estudiantes de una manera correcta. Ahora, si bien no hay un valor exacto para la cantidad mínima para alcanzar dicha meta, en algunas implementaciones se habla de por lo menos tres veces el tamaño esperado de una prueba tradicional (Fasttest, 2013).

Respecto a los elementos segundo y tercero, diversos trabajos como e-Adaptive (Kustiyahningsih & Dwi Cahyani, 2013), Flip (Barla *et al.*, 2010), CIA (Jiménez *et al.*, 2008), AHA 3.0 (De Bra *et al.*, 2007), SIETTE (Conejo *et al.*, 2004), e Inspire (Papanikolaou *et al.*, 2003), los definen a partir de la Teoría de Respuesta al Ítem o TRI. Esta aproximación brinda bases probabilísticas al problema de la medición de rasgos indirectamente observables, también denominados rasgos latentes, como lo es el nivel de conocimiento. Su nombre se debe a la consideración de los ítems (preguntas) como unidades fundamentales de las evaluaciones, y no tanto la puntuación final, como si lo hacen las aproximaciones tradicionales (Lord, 1980; Hambleton *et al.*, 1991; Hambleton & Jones, 1993; Gil & Suárez, 2003; Matas *et al.*, 2004; Chang & Ying, 2009; Liu *et al.*, 2010; Muñiz, 2010; Wauters *et al.*, 2010).
Según esta teoría, la relación entre el rasgo θ, entendido como el nivel de conocimiento del evaluado en un determinado tema, y su respuesta a cada ítem de evaluación, puede ser explicado mediante una función monótona creciente denominada Curva Característica del Ítem, o CCI, la cual establece la probabilidad de acierto (Traub & Wolfe, 1981). Dependiendo de la naturaleza y parámetros de dicha función existen diferentes modelos a ser usados, siendo algunos de los más populares los siguientes.

- El modelo Rash, también conocido como 1PL, caracterizado por presentar una forma logística y contar con un único parámetro: la dificultad del ítem.
- La ojiva, sea normal o logística, con dos parámetros: dificultad y discriminación del ítem. La versión logística, también conocida como 2PL es la más empleada de las dos.
- La ojiva, sea normal o logística, con tres parámetros: dificultad b, discriminación a y adivinanza c del ítem. Similar al caso anterior, la versión logística, también conocida como 3PL es la más empleada de las dos.

De los tres, el 3PL es el más general. De hecho, los modelos 1Pl y 2PL pueden ser concebidos como casos particulares, es decir con parámetros constantes del 3PL. La fórmula matemática de dicho modelo puede ser expresada como:

$$P(\theta) = c + \frac{1-c}{1+e^{-a(\theta-b)}} \tag{1}$$

Con el fin de ilustrar la fórmula anterior, así como para clarificar el papel de los parámetros involucrados, la Figura 1 presenta una curva 3PL típica. Pese a que el rango de dicha función en términos de θ es $(-\infty, \infty)$, basta con considerar el intervalo cerrado $[-3, 3]$ para fines prácticos

dada la forma asintótica de la función. Entre tanto, el dominio de la función es el intervalo abierto $(c, 1)$ donde ambos extremos representan sus límites asintóticos (Harris, 1989).

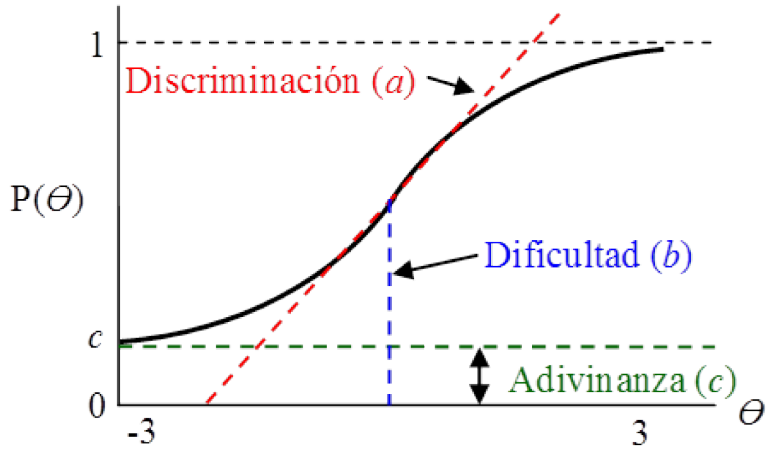

Figura 1: CCI típica del modelo 3PL. Fuente: Elaboración propia.

Respecto a los parámetros, y en el contexto de la TRI, la adivinanza c define la probabilidad de acierto intrínseca del ítem. En otras palabras, este parámetro es inherente a la forma de presentación del ítem. Así por ejemplo, en una pregunta de tipo falso/verdadero, si el evaluado no sabe la respuesta correcta, existe una probabilidad de 0.5 de que aun así acierte. De forma contraria, el parámetro de dificultad b, como su nombre lo indica, se refiere al grado de dificultad del ítem independiente de cómo es presentado. Así por ejemplo, una pregunta en el contexto de química para educación secundaria como "¿cuál es el peso atómico del azufre?" puede considerarse difícil bien sea que se pregunte como respuesta libre, opción múltiple o incluso falso/verdadero. En términos gráficos, este parámetro este parámetro define que tan a la derecha de la curva el ítem "desafía" a un evaluado con un nivel de conocimiento alto o, de manera recíproca, que tan a la izquierda predispone un acierto de un evaluado con un nivel de conocimiento bajo. En palabras más técnicas, este parámetro determina la ubicación del punto de inflexión de la curva a lo largo del eje de θ. Finalmente, el parámetro de discriminación a, como su nombre lo sugiere, determina que tan progresivo a lo largo del eje de θ es el ítem para discriminar las posibilidades de acierto y de fallo. En términos gráficos este parámetro determina el ángulo de inclinación de la curva en su punto de inflexión. Esto quiere decir que, a más inclinada la curva, más angosto será el umbral de θ donde la probabilidad de acierto o fallo no esté bien definida. Por el contrario, a más aplanada la curva, la probabilidad de acierto será menos abrupta a lo largo de θ. Dicho de otra manera, mientras más plana sea la curva, menor será la capacidad del ítem para discriminar entre dos estudiantes con niveles de conocimiento estimado similares y viceversa.

Para esclarecer aún más el impacto de estos dos últimos parámetros sobre la CCI, la figura 2 muestra diferentes curvas a partir de la variación de dichos parámetros dejando los demás constantes. Nótese

que el valor de c es cero en estos casos.

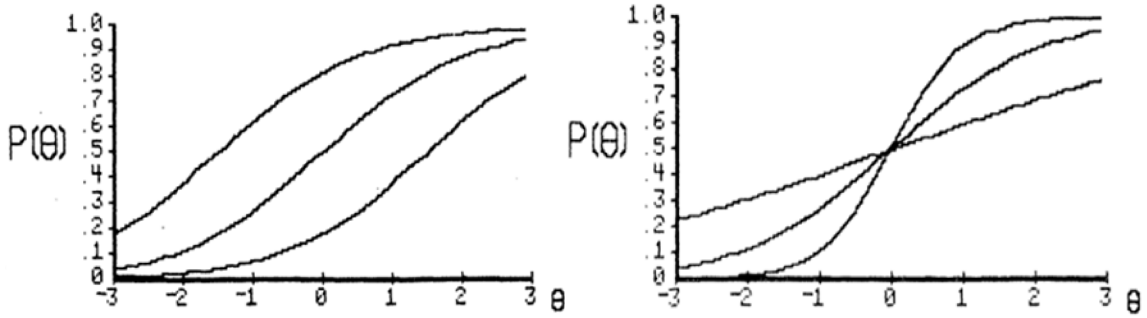

Figura 2: CCI del modelo 3PL variando el parámetro b (izquierda) y variando el parámetro a (derecha). Fuente: Baker (2001)

En el modelo propuesto se siguieron los siguientes lineamientos para determinar los valores de estos tres parámetros para un ítem. Primero, considerando que c depende del tipo de pregunta, es decir el formato en que es presentado, y no lo la pregunta en sí, este valor puede ser calculado automáticamente como se muestra en la tabla 1 para algunos de los tipos de preguntas más comunes.

Tabla 1: Parámetro de adivinanza según el tipo de pregunta. Fuente: Elaboración propia.

Tipo de pregunta	**Consideraciones**	**Valor del parámetro** c
Falso/Verdadero	Ninguna	0.5
Opción múltiple con respuesta única	n: cantidad de opciones	n^{-1}
Opción múltiple con respuesta múltiple	n: cantidad de opciones	$\left[1+\sum_{i=1}^{n-1}\frac{n!}{i!(n-i)!}\right]^{-1}$
Emparejamiento	nA: cantidad de elementos ordenados nB: cantidad de elementos desordenados ($nB \geq nA$)	$\left[\frac{n_B!}{(n_B-n_A)!}\right]^{-1}$
Ordenamiento	n: cantidad de elementos a ordenar	$(n!)^{-1}$
Respuesta libre	ninguna	0

Segundo, para el parámetro de dificultad se propone que, en caso de no contar con un registro histórico de evaluaciones, la determinación de dicho parámetro, dentro del rango $(-3, 3)$ y de forma lineal, corra por cuenta del docente con base a su experiencia. Ahora, en caso que se cuente con un registro histórico, dicho valor podría estimarse a partir del porcentaje de aciertos de los sujetos de prueba descontando previamente el factor de adivinanza. Es importante aclarar sin embargo que dicha calibración requeriría de un número considerable de sujetos, no menos de 1000 (Kozierkiewicz-Hetmańska, A. & Poniatowski, 2014; Wainer & Mislevy, 2000).

Tercero, y de manera similar a con el parámetro anterior, se propone para el caso de no contar con registros históricos utilizar un valor constante de 1.0 para el parámetro de discriminación, el cual

se traduce en una CCI con una pendiente neutral (de 45°). Igualmente, al momento de contar con registros históricos dicho valor podría calibrarse haciendo el mapeo con los niveles de conocimiento estimados de los evaluados en el momento que presentaron cada ítem.

Una vez definida la forma de la CCI, es necesario precisar cómo se realiza la selección de ítems y cómo se estima el nivel de conocimiento del evaluado. Así mismo, es necesario precisar el cuarto elemento de la EAC: el criterio de terminación de la evaluación. Para selección de ítems se hace uso de lo que se conoce como Función de Información del Ítem o FFI, que se calcula a partir de la correspondiente CCI. Para el caso del modelo 3PL dicha función toma la siguiente forma (Barla *et al.*, 2010):

$$I_i(\theta) = a^2 \left[\frac{Q_i(\theta)}{P_i(\theta)}\right]\left[\frac{P_i(\theta) - c_i}{1 - c_i}\right]^2 \tag{2}$$

En este caso, si $P_i(\theta)$ representa la probabilidad de acierto, $Q_i(\theta)$ representa la probabilidad de falla. De esta manera, dado un estimado de θ para un evaluado, el ítem más apropiado dentro del repositorio disponible, de tamaño V, se obtiene mediante la siguiente fórmula:

$$max_i \{I_i(\theta)\} \quad \text{for} \quad i = 1 \quad \text{to} \quad V \tag{3}$$

Por su parte, la manera más común de estimar el nivel de conocimiento del evaluado se basa en la función de máxima verosimilitud que consiste básicamente en encontrar el valor de θ que maximiza la función definida por la siguiente fórmula.

$$L(u|\theta) = \prod_{i=1}^{n} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i} \tag{4}$$

Donde Q_i es la probabilidad de no acertar en el ítem $i(1 - P_i)$ mientras que $u = (u_1, ..., u_W)$ es el vector de respuestas del evaluado. Para $i = 1, ..., W$, u_i es 1 si la respuesta al ítem i es correcta y 0 en caso contrario. Luego, se emplea el método a priori (Baker, 2001) que se basa en los valores de los parámetros de los ítems y, como su nombre lo indica, en el valor previamente estimado del nivel de conocimiento del evaluado. Más específicamente, hace uso de un procedimiento iterativo según la siguiente fórmula.

$$\theta_{s+1} = \theta_s + \frac{\sum_{i=1}^{W} a_i (1 - c_i)(u_i - P_i(\theta_s))}{\sum_{i=1}^{W} I_i(\theta_s)} \tag{5}$$

Por último, como criterio de terminación de la evaluación se establece que dejan de presentarse ítems al evaluado cuando ocurra por lo menos una de dos condiciones: cuando la estimación del nivel de conocimiento supere un valor de 2.95, o cuando la cantidad de ítems presentados supere un umbral máximo definido por el docente. El valor de 2.95 se debe a la forma asintótica de la CCI con valores extremos en -3 y 3. Un esquema del modelo completo presentado en esta sección se presenta en la figura 3.

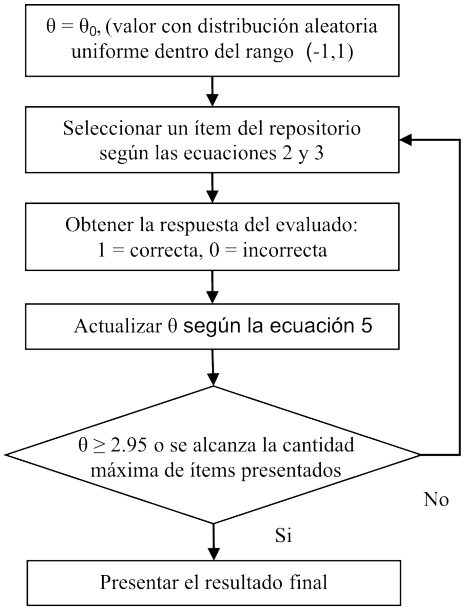

Figura 3: Esquema general del modelo de evaluación. Fuente: Elaboración propia.

3. METODOLOGÍA DE VALIDACIÓN Y RESULTADOS

Con el fin de validar el modelo presentado se llevó a cabo un estudio en la Universidad Nacional de Colombia dentro del curso Estructura de datos ofrecido dentro del programa de Ingeniería de Sistemas e Informática. La muestra estuvo compuesta por 59 estudiantes de primer año, 51 hombres y 8 mujeres, con edad promedio de 17.51 y desviación estándar de 1.68. Dicha muestra fue aleatoriamente dividida en dos grupos, uno de control con 30 estudiantes y otro experimental con los restantes 29.

Con el grupo de control se llevó a cabo una evaluación con 10 preguntas: dos de tipo falso/verdadero, dos de opción múltiple con respuesta única, dos de opción múltiple con respuesta múltiple, dos de ordenamiento y dos de respuesta libre. La cantidad de preguntas y la formulación de las mismas se determinó de tal manera que el desarrollo de la prueba resultante se ajustara a los parámetros de una prueba normal dentro del curso.

Esta evaluación se implementó mediante la plataforma Moodle y la única diferencia entre estudiantes fue el orden de presentación de las preguntas, el cual fue aleatorio. Entre tanto para el grupo experimental se utilizó un repositorio con 20 preguntas, con el doble de preguntas por cada tipo respecto al grupo de control. En este caso la evaluación se implementó mediante una plataforma particular desarrollada al interior de la universidad que empleó el modelo propuesto. Dicha plataforma exhibe una interfaz y un comportamiento similar a Moodle, solo que incorpora varias funcionalidades adaptativas, entre ellas la evaluación. Cabe señalar sin embargo que solo esta última fue utilizada para garantizar que ambas poblaciones fueran comparables.

Tanto en el banco de preguntas del grupo de control como en el del experimental las preguntas tenían diferentes niveles de dificultad b, especificadas en todos los casos por el docente responsable. Por su parte, el factor de adivinanza de cada pregunta c se calculó a partir de las fórmulas presentadas en la Tabla 1, mientras que para el factor de discriminación a se usó en todos los casos un valor de 1.0.

El objeto del estudio era determinar si existía una diferencia significativa entre los resultados de la evaluación realizada de modo tradicional (grupo de control), es decir con un número fijo de preguntas, comparada con el modelo propuesto (grupo experimental). Dicha diferencia se consideró bajo dos aspectos: por un lado la precisión de la evaluación y por otro el tiempo necesario para realizarla.

Para el primer aspecto se tomó como punto de referencia la calificación consolidada de los estudiantes de ambos grupos en un curso previo sobre el tema específico objeto de la evaluación. Dicho tema fueron conocimientos básicos en lógica de programación. Todos los 59 estudiantes tomaron dicho curso bajo las mismas condiciones y en el mismo periodo de tiempo. Dicha calificación, al menos en teoría, se consideró como la mejor estimación del nivel de conocimiento real de los estudiantes pues se trató de una evaluación continua a lo largo de todo un semestre.

En la figura 4 se muestran los datos correspondientes considerando que las calificaciones se encuentran en una escala de 0 a 100 sin cifras decimales. El eje de las abscisas corresponde a los estudiantes y el eje de las ordenadas a las calificaciones. Para facilitar la visualización, los datos se muestran en orden ascendente.

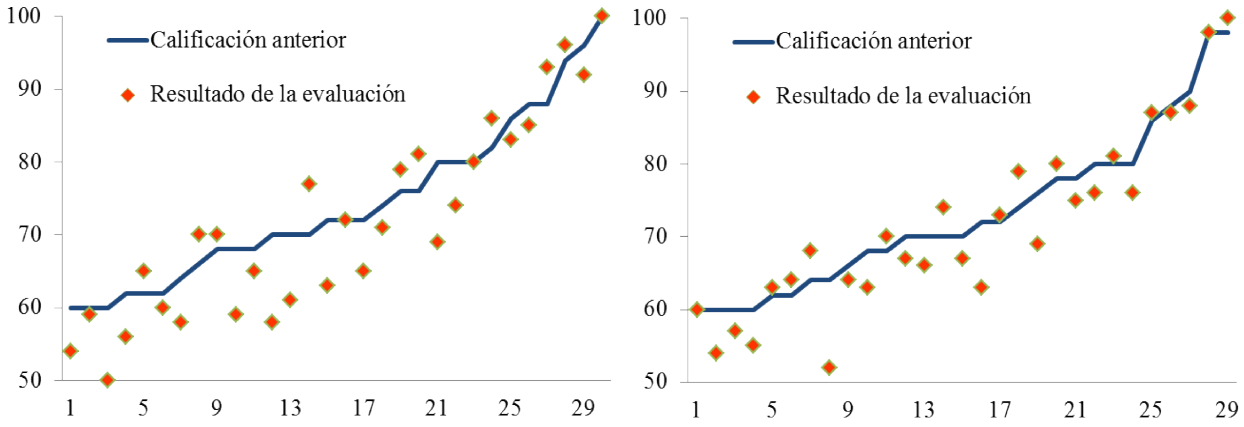

Figura 4: Resultados para el grupo de control (izquierda) y el experimental (derecha). Fuente: Elaboración propia.

Se realizaron dos análisis estadísticos para comparar el proceso de evaluación en ambos grupos, específicamente si los resultados obtenidos concordaban con la nota previa de los estudiantes. El primer análisis fue una prueba para muestras independientes por medio del estadístico t. En este

caso se usó como hipótesis nula H_0 que las medias eran iguales, contra la hipótesis alterna H_1 que no lo eran. El segundo análisis fue una prueba de tamaño de efecto por medio del estadístico d de Cohen. En ambos casos los datos presentados en la tabla 2 fueron usados.

Tabla 2: Resumen de calificaciones previas y resultados de evaluación. Fuente: Elaboración propia.

	Calificación previa			**Resultados de evaluación**	
Grupo	**Número de estudiantes**	**Media**	**Desviación estándar**	**Media**	**Desviación estándar**
Control	30	74.20	11.18	71.70	13.41
Experimental	29	73.24	10.95	71.59	12.35

El valor del estadístico t para el grupo de control fue 0.78 con un correspondiente valor P de 0.44. Como este valor es mayor a 0.05 se puede decir, con un nivel de significancia del 95 %, que se acepta la hipótesis nula. En otras palabras no existe evidencia estadística de la diferencia en la media de la calificación anterior comparada con los resultados de la evaluación llevada a cabo.

Consecuentemente con este resultado, el valor del estadístico d de Cohen fue de 0.21 lo cual implica un tamaño de efecto muy pequeño. Cabe resaltar que el tamaño del efecto medido a través de este estadístico es una medida de la fuerza de un fenómeno, en este caso la diferencia de resultados entre dos grupos dentro de un experimento. Dicho tamaño sirve como una estadística descriptiva para complementar la estadística inferencial, como son los valores P. Se considera que valores inferiores a 0.2 en valor absoluto indican un efecto de pequeño tamaño, cercanos a 0.5 indican una magnitud media, y valores por encima de 0.8 indican un efecto de alta magnitud (Cohen, 1988).

Un panorama similar se presenta en el caso del grupo experimental donde el valor del estadístico t fue 0.54 con un correspondiente valor P de 0.59, mientras que el estadístico d de Cohen fue de 0.14, lo cual conlleva a la misma conclusión.

Un método adicional para determinar si hubo o no diferencia entre las medidas es usando un análisis de residuales vía raíz del error medio cuadrado, más conocido como RMSE por sus siglas en inglés. Aquí, el término residual o error es usado en el sentido de la desviación respecto al valor esperado, en este caso, de la calificación previa. La figura 5 presenta los correspondientes diagramas de frecuencia de los residuales.

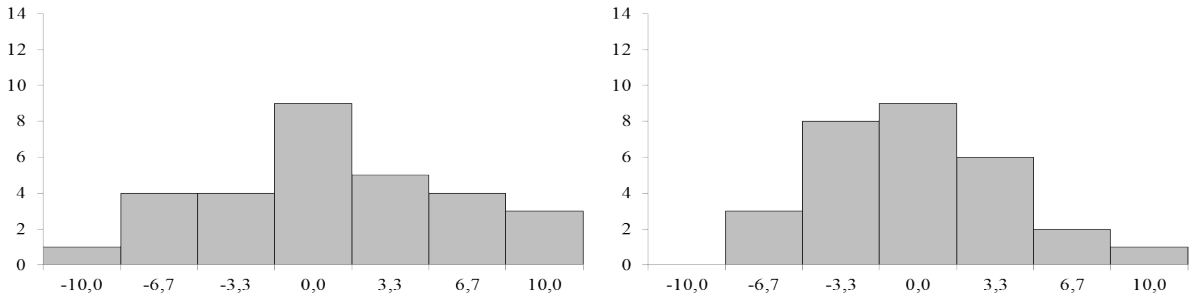

Figura 5: Diagrama de frecuencia de los residuales para el grupo de control (izquierda) y para el experimental (derecha). Fuente: Elaboración propia.

El valor del RMSE del grupo de control fue de 5.88 y del grupo experimental 4.41. Este resultado, junto con los dos análisis previamente descritos, demuestra que tanto la evaluación realizada mediante un método tradicional como la del método propuesto brindaron mediciones acertadas del nivel de conocimiento de los estudiantes según los datos que se tenían previamente. Existe sin embargo una pequeña diferencia a favor del modelo propuesto según lo evidencian los valores P, estadísticos d de Cohen y RMSE. Dicha diferencia, según se puede observar en la figura 4, es casi imperceptible en los estudiantes con notas más altas pero es mucho más notoria con el resto.

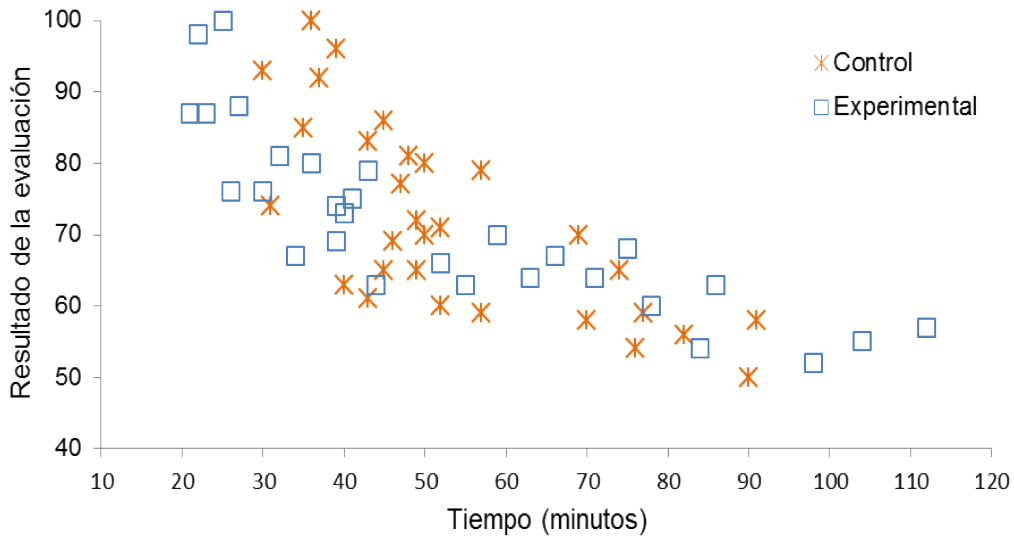

Figura 6: Tiempos de evaluación versus resultado de la evaluación del grupo de control y experimental. Fuente: Elaboración propia.

En ambos grupos el tiempo de evaluación fue, en términos generales, inversamente proporcional a las calificaciones obtenidas. La tabla 3 muestra un resumen de estos datos y, para efectos de comparación, los estudiantes de ambos grupos se muestran en general pero también divididos en tres poblaciones: calificaciones bajas (inferiores a 74), calificaciones medias (entre 74 y 86, inclusive),

y calificaciones altas (superiores a 86). Dichos intervalos fueron elegidos dividiendo el rango de las calificaciones obtenidas en tres igualmente espaciados.

Tabla 3: Resumen de los tiempos de evaluación. Fuente: Elaboración propia.

Población de estudiantes	**Grupo**	**Media**	**Desviación estándar**	**Estadístico t**	**Valor P**	**d de Cohen**
General	Control	53.67	17.12	0.19	0.85	0.05
	Experimental	52.59	26.37			
Calificaciones bajas	Control	61.78	16.80	1.23	0.22	-0.32
	Experimental	68.24	23.40			
Calificaciones medias	Control	44.50	8.28	4.88	0.000	1.28
	Experimental	35.29	6.21			
Calificaciones altas	Control	35.50	3.87	14.29	0.000	3.74
	Experimental	23.60	2.41			

Igual que en el caso de comparación de calificaciones, se realizó una prueba de hipótesis así como un análisis de tamaño de efecto para los tiempos de evaluación. Esta vez se usó como hipótesis nula H_0 que las medias del tiempo eran iguales en ambos grupos, contra la hipótesis alterna H_1 que no lo eran. Como puede observarse en la tabla 3 al comparar las poblaciones generales de estudiantes el valor P es mayor que 0.05 por lo cual se acepta la hipótesis nula. En otras palabras, no existe evidencia estadística significativa de la diferencia en el tiempo de evaluación de los dos grupos. Consecuentemente, el estadístico d de Cohen implica un tamaño de efecto muy bajo.

Algo similar ocurre al comparar las poblaciones de calificaciones bajas, las cuales por cierto son las que presentan la mayor desviación estándar en ambos casos. Sin embargo, un panorama diferente ocurre con las poblaciones de calificaciones medias y bajas. En estos casos los valores P son inferiores a 0.001 lo cual indica que se rechaza la hipótesis nula en favor de la alterna. Lo que esto significa es que hay una evidencia estadística, con un nivel de significancia del 99 % de una diferencia entre los tiempos medios de evaluación. De hecho, dicha diferencia es en promedio de casi 9.2 minutos a favor del grupo experimental en el caso de la población de calificaciones medias, y de casi 11.9 minutos en el caso de la población de calificaciones altas. Consecuentemente, los respectivos valores del estadístico d de Cohen indican un tamaño de efecto alto.

Con el fin de reforzar estos hallazgos se llevó a cabo una regresión lineal simple para ambos grupos usando el tiempo como variable dependiente y la calificación obtenida como variable independiente. En el caso del grupo de control la regresión fue $-0{,}93x + 120{,}16$ con un R^2 de 0.53 mientras para el grupo experimental fue $-1{,}78x + 180{,}08$ con un R^2 de 0.7. Como puede notarse la pendiente del grupo de control es considerablemente más plana que la del experimental.

Lo que estos resultados indican es que en promedio un estudiante, en particular aquellos con un nivel de conocimiento medio o alto, requiere de considerablemente menos tiempo para realizar una evaluación si se realiza mediante el modelo propuesto que si se realiza de forma tradicional.

4. CONCLUSIONES

La evaluación de estudiantes por medio de pruebas computarizadas se ha vuelto con los años una práctica cada vez más común y no solo en cursos enteramente virtuales sino incluso como complemento en cursos presenciales. Esta 'popularidad' se debe en una parte a la automatización que este tipo de pruebas implica, la cual se traduce en velocidad, precisión y escalabilidad. Y las ventajas no terminan allí, también permiten eliminar la subjetividad de un evaluador humano y brindan la capacidad de brindar una realimentación inmediata.

Existe sin embargo una crítica recurrente hacia esta aproximación: todos los estudiantes son tratados de la misma manera (pese a la aleatoriedad que pueda emplearse) produciendo muchas veces que los estudiantes 'avanzados' se sientan poco estimulados o incluso aburridos, al tiempo que estudiantes 'rezagados' se sientan frustrados o intimidados. Una alternativa para solucionar este problema se conoce como Evaluación Adaptativa Computarizada, la cual se basa en una construcción dinámica de evaluaciones a partir de la estimación del nivel de conocimiento del estudiante.

Considerando este escenario, la investigación descrita en este artículo presenta una contribución en el área proponiendo y validando un modelo basado en la Teoría de Respuesta al Ítem. Para el proceso validación se llevó a cabo un estudio con estudiantes de ingeniería en una universidad Colombiana el cual demostró no solo que el modelo provee en promedio un mayor nivel de precisión en la medición del nivel de conocimiento sino también un tiempo menor de evaluación contrastado con el método tradicional.

Respecto al primer aspecto, la precisión, es importante mencionar que para los estudiantes con mayor nivel de conocimiento resulta prácticamente independiente la forma de evaluación, en cuanto que para los estudiantes de niveles medio y bajo el modelo propuesto exhibe un mayor ajuste así como que una menor dispersión.

Respecto al segundo aspecto, el tiempo, ocurre un fenómeno similar pero con la población contraria. Al analizar las poblaciones de estudiantes discriminados según niveles de conocimiento: bajo, medio, y alto; resultó que para los primeros no existe una evidencia estadística de una diferencia en el tiempo promedio que les lleva hacer la evaluación. Entre tanto para los estudiantes de conocimiento medio y alto se percibe una significativa disminución en dicho tiempo cuando se emplea el modelo propuesto.

A pesar de estos hallazgos, es importante resaltar que también existen desventajas en el enfoque presentado. Una de ellas es el esfuerzo adicional que requiere la creación del repositorio de ítems. Lo anterior considerando, no solo su definición como tal, sino la determinación y calibración de parámetros. Esto en términos prácticos implica una mayor dedicación por parte de los docentes a la hora de adoptar el enfoque.

Otro punto a resaltar es que el experimento realizado solo puede aportar conclusiones preliminares respecto al modelo empleado. Por una parte porque solo se realizó sobre un grupo de estudiantes de un tamaño reducido, y por otra porque solo se llevó a cabo para una evaluación. Precisamente, como trabajo futuro, se espera realizar futuros experimentos que contemplen tales aspectos.

Referencias

Baker, F. (2001), The Basics of Item Response Theory, Second edition. ERIC Clearinghouse on Assessment and Evaluation, 172.

Barla, M.; Bieliková, M.; Ezzeddinne, A.; Kramár, T.; Simko, M. & Vozár, O. (2010), On the impact of adaptive test question selection for learning efficiency. *Computers & Education*, 55(**2**), 846–857.

Chang, H-H. (2014), Psychometrics behind Computerized Adaptive Testing. *Psychometrika*, 79, 1-20.

Chang, H-H. & Ying, Z. (2009), Nonlinear Sequential Designs for Logistic Item Response Theory Models with Applications to Computerized Adaptive Tests. *The Annals of Statistics*, 37(**3**), 1466-1488.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (second ed.), Lawrence Erlbaum Associates.

Conejo, R.; Guzmán, R.; Millán, E.; Trella, M.; Pérez-de-la-cruz, J. & Ríos, A. (2004), SIETTE: a web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education*, 14(**1**), 29–61.

De Bra, P.; Stash, N.; Smits, D.; Romero, C. & Ventura, S. (2007), Authoring and Management Tools for Adaptive Educational Hypermedia Systems: the AHA! case study. *Studies in Computational Intelligence*, 62, 285-308.

Eggen, T. & Straetmans, G. (2000), Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement.* 60(**5**), 713-734.

Fasttest (2013), *Requirements of Computerized Adaptive Testing.* Disponible en: http://www.fasttestweb.com/ftw-docs/CAT_Requirements.pdf.

Gil, G. & Suárez, J. (2003), Sistemas de presentación de los resultados de las evaluaciones del rendimiento educativo: aplicación al estudio internacional de la lengua inglesa en la educación secundaria. *Revista de Investigación Educativa*, 21(**1**), 135-155.

Guzmán, E.; Conejo, R. & Pérez de la Cruz, J. (2007), Adaptive testing for hierarchical student models. User Modeling and User-Adapted Interaction. 17 (**1-2**), 119-157.

Hambleton, R. & Jones, R. (1993), Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practices*, 12(**3**), 38-47.

Hambleton, R.; Swaminathan, H. & Rogers, H. (1991), Fundamentals of Item Response Theory. Newbury Park, California: Sage publications, 184.

Harris, D. (1989), Comparison- of 1-, 2-, and 3-Parameter IRT *Models.Educational Measurement Issues and Practice*, 8(**1**), 35-41.

Huang, S. (1996), A Content-Balanced Adaptive Testing Algorithm for Computer-Based Training Systems. Third International Conference in Intelligent Tutoring Systems. Montreal, 306-314.

Jiménez, M.; Jiménez, J. & Ovalle, D. (2008), Un modelo de evaluación del conocimiento para cursos virtuales adaptativos usando la teoría de respuesta al ítem. In: Tendencias en Ingeniería de Software e Inteligencia Artificial, 2, 91-98.

Kozierkiewicz-Hetmańska, A. & Poniatowski, R. (2014), An Item Bank Calibration Method for a Computer Adaptive Test. *Lecture Notes in Computer Science*, 8397, 375-383.

Kustiyahningsih, Y. & Dwi Cahyani, A. (2013), Computerized Adaptive Test based on Item Response Theory in E-Learning System. *International Journal of Computer Applications*, 81(**6**), 6-11.

López, R.; Sanmartín, P. & Méndez, F. (2014), Revisión de las evaluaciones adaptativas computarizadas (CAT), *Educación y humanismo*, 16(**26**), 27-40.

Lord, F. (1980), Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 280.

Liu, C.; Meng, P.; Zhang, Z. & Pan, Y. (2010), Research on computerized adaptive testing system based on IRT. *Proceedings of the Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, 4, 1817-1821.

Matas, A.; Tójar, J.; Jaime, J.; Benítez, F. & Almeda, L. (2004), Diagnóstico de las actitudes hacia el medio ambiente en alumnos de secundaria: una aplicación de la TRI. *Revista de Investigación Educativa*, 22(**1**), 233-244.

Ministerio de Educación Nacional de Colombia (2003), Los beneficios de evaluar. Altablero, 19, 9. Dispoible en: http://www.mineducacion.gov.co/1621/propertyvalue-31340.html

Muñiz, J. (2010), Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo*, 31(**1**), 57-66.

Oppermann, R. & Kinshuk, R. (1997), Adaptability and Adaptivity in Learning Systems. *Knowledge Transfer*, 2, 173-179.

*Ö*zyurt, H.; *Ö*zyurt; *Ö*., Baki, A. & Güven, B. (2012), Integrating computerized adaptive testing into UZWEBMAT: Implementation of individualized assessment module in an e-learning system. *Expert Systems with Applications*, 39(**10**), 9837-9847.

Papanikolaou, K. A.; Grigoriadou, M.; Kornikolakis, H. & Magoulas, G. D. (2003), Personalizing the interaction in a web-based educational hypermedia system: the case of Inspire. *User Modeling and User-Adapted Interaction*, 13(**3**), 213-267.

Ponsoda, V. (2000), Overview of the Computerized Adaptive Testing. *Psicológica*, 21(**1**), 115-120.

Thissen, D. & Mislevy, R. (2000), Testing Algorithms. In Wainer, H. (Ed.) *Computerized Adaptive Testing: A Primer*, 101-134. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Thompson, N. & Weiss, D. (2011), A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research & Evaluation*, 16(**1**), 1-9.

Traub, R. & Wolfe, R. (1981), Latent Trait Theories and the Assessment of Educational Achievement. *Review of Research in Education*, 9, 377-435.

Trevitt, C.; Brenan, E. & Stocks, C. (2012), Evaluación y aprendizaje: ¿Es ya el momento de replantearse las actividades del alumnado y los roles académicos? *Revista de Investigación Educativa*, 30(**2**), 253-267.

Triantafillou, E.; Georgiadou, E. & Economides, A. (2008), The design and evaluation of a computerized adaptive test on mobile devices. *Computers & Education*, 50(**4**), 1319-1330.

Van der Linden, W. & Glas, C. (2010), *Elements of Adaptive Testing*. New York: Springer. 31-56.

Wainer, H. & Mislevy, R. (2000), Item response theory, calibration, and estimation. En: Wainer, H. (ed.) *Computerized Adaptive Testing: A Primer*. Mahwah: Lawrence Erlbaum Associates.

Wauters, K.; Desmet, P. & Van den Noortgate, W. (2010), Adaptive item-based learning environments based on the item response theory: possibilities and challenges. *Journal of Computer Assisted Learning*, 26(**6**), 549-562.

Weiss, D. & Kingsbury, G. (1984), Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*. 21, 361-375.