Application of a Computerized Adaptive Test in Statics Course

Jorge Luis Restrepo Ochoa, Ph.D[1], Jaime Leonardo Barbosa Pérez, M.Sc[2], and Julian Arenas Berrio, Ing.[3]
[1,2,3]Universidad EAFIT, Colombia, jrestrep@eafit.edu.co, jbarbosa@eafit.edu.co, jarenas5@eafit.edu.co

***Abstract*– In the virtual platform for Statics course of the Engineering School at EAFIT University a Computer Based Test (CBT) and a Computerized Adaptive Test (CAT) were applied to measure the student's ability in previous subjects required for the Statics course. It was proved that the CAT's tests determine the student's ability with less amount of questions and take the examinee to the threshold of his capacity trough increasing levels of difficulty in contradiction to CBT's tests in which a fixed set of items is used. But in CAT's the duration time for the execution is higher. This paper presents the analysis of the collected data in both tests and it exposes the conclusions on the CAT's tests for the education in engineering.**

***Keywords*— Evaluation of Learning, Adaptability, Adaptive Tests, Computerized Tests, b-learning.**

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2016.1.1.280
ISBN: 978-0-9822896-9-3
ISSN: 2414-6390

Aplicación de una prueba adaptativa computarizada en la asignatura Estática

Jorge Luis Restrepo Ochoa, Ph.D[1], Jaime Leonardo Barbosa Pérez, M.Sc[2], and Julian Arenas Berrio, Ing.[3]
[1,2,3]Universidad EAFIT, Colombia, jrestrep@eafit.edu.co, jbarbosa@eafit.edu.co, jarenas5@eafit.edu.co

***Abstract*– In the virtual platform for Statics course of the Engineering School at EAFIT University a Computer Based Test (CBT) and a Computerized Adaptive Test (CAT) were applied to measure the student's ability in previous subjects required for the Statics course. It was proved that the CAT's tests determine the student's ability with less amount of questions and take the examinee to the threshold of his capacity trough increasing levels of difficulty in contradiction to CBT's tests in which a fixed set of items is used. But in CAT's the duration time for the execution is higher. This paper presents the analysis of the collected data in both tests and it exposes the conclusions on the CAT's tests for the education in engineering.**

***Keywords*— Evaluation of Learning, Adaptability, Adaptive Tests, Computerized Tests, b-learning.**

***Resumen*– En la plataforma virtual para la asignatura Estática de la Escuela de Ingeniería de la Universidad EAFIT se aplicó una prueba basada en computador (CBT) y una prueba adaptativa computarizada (CAT) para medir la habilidad de los estudiantes en temas requeridos para el curso. Se comprobó que las pruebas CAT determinan la habilidad del estudiante en menor número de preguntas y llevan al examinado hasta el umbral de su capacidad a través de items con niveles de dificultad variables de acuerdo con el desempeño de cada estudiante, a diferencia de las pruebas CBT en las que se utiliza un conjunto de items fijo para todos los estudiantes. Sin embargo en las pruebas CAT el tiempo de aplicación fue mayor. Este trabajo presenta el análisis de los datos recolectados en ambas pruebas y expone las conclusiones sobre la prueba CAT para enseñanza en Estática.**

***Palabras clave*—Evaluación del aprendizaje, Adaptabilidad, Pruebas adaptativas, Pruebas computarizadas, b-learning**

I. Introducción

Debido al bajo rendimiento en la asignatura Estática para ingeniería en la universidad EAFIT, cuyo índice de reprobación de estudiantes matriculados era de 48.8% y la deserción se encontraba alrededor del 30% en los registros oficiales hasta 2011, la Escuela de Ingeniería implementó una plataforma virtual para mejorar los procesos de aprendizaje de los estudiantes y, por consiguiente su rendimiento académico [1] [2] [3].

La investigación alrededor de la plataforma virtual ha estado direccionada a la determinación de los parámetros característicos de dificultad y discriminación de las preguntas, y a su vez, la fiabilidad de los entrenamientos y pruebas suministradas a los estudiantes mediante la teoría de respuesta al ítem o IRT (Item Response Theory), con el objetivo de mejorar la calidad de los contenidos y del sistema [4] [5] [6].

Ahora bien, según la investigación sobre tendencias en la educación mundial, NMC HorizonReport: 2014 Higher Education Edition [7], para los próximos años se espera que la educación se adapte a las expectativas de los estudiantes y sus necesidades, tanto en el proceso de aprendizaje como en la evaluación de conocimientos. Actualmente en la asignatura Estática se aplican pruebas basadas en computador o CBT (Computerized based tests), las cuales se presentan de forma lineal y preestablecida al estudiante, pero con el fin de ajustar el proceso de evaluación con las tendencias mencionadas,en el presente estudio se aplica una prueba adaptativa computarizada o CAT (Computerized Adaptive Test) cuyas preguntas varían de acuerdo al desempeño de cada estudiante.

En la literatura se encuentran casos de implementación de plataformas virtuales para la enseñanza de Estática en ingeniería [8][9], pero no se encuentran registros con evaluaciones adaptativas para dicha asignatura. Las pruebas CAT, son utilizadas generalmente en contextos clínicos, exámenes estandarizados (e.g. GMAT, GRE, MCSE, TOEFL) y evaluación del desempeño en idiomas [10]. Para la evaluación del aprendizaje en ciencias básicas mediante pruebas tipo CAT, un punto de referencia es el sistema SIETTE [11] [12], desarrollado en la universidad de Málaga y extendido a otras universidades en España, no obstante no se muestra un estudio específico en asignaturas básicas en ingeniería como es Estática, el cual se aborda en esta investigación.

Al inicio del semestre 2016-1, se aplicó a 118 estudiantes de ingeniería civil, mecánica y de producción que cursan la asignatura Estática, dos pruebas para medir sus conocimientos en conversión de unidades, geometría, trigonometría y solución de ecuaciones, temas que los docentes identificaron como requeridos para lograr un buen desempeño en la asignatura. La primera prueba es adaptativa cuyos requisitos de aplicación Weiss [13] los enuncia de la siguiente forma: Banco precalibrado de preguntas, nivel de dificultad de la pregunta inicial, algoritmo iterativo de selección de la siguiente pregunta basado en la respuestas anteriores y un criterio de parada o terminación. La segunda prueba es de tipo lineal con ítems fijos aunque se varan los valores para cada estudiante con el fin de incrementar la seguridad de la prueba.

El objeto del estudio es comprobar que las pruebas CAT determinan la habilidad del estudiante con menor número de preguntas, en menor tiempo y con niveles de dificultad que llevan al examinado hasta el umbral de su capacidad tal como lo afirman diversos autores [10] [13] [14].

II. MATERIALES Y MÉTODO

A. *Arquitectura del sistema*

La plataforma virtual se encuentra implementada en el sistema de gestión de contenidos de aprendizaje MOODLE, con licencia libre GNU (General public license) y albergada en los servicios de computación en la nube de AWS Amazon.

Para la generación de preguntas en formato compatible con MOODLE, se cuenta con una aplicación de escritorio desarrollada en lenguaje JAVA, cuyo propósito es programar el algoritmo de solución para cada pregunta y variar en un rango definido los datos de las variables relacionadas en el ejercicio. Mediante un estudio anterior se comprobó que variar los datos no modifica la dificultad de la pregunta [5], y en consecuencia, no cambia sustancialmente la habilidad a evaluar. Las preguntas utilizadas tanto en pruebas CBT como en pruebas CAT son de opción múltiple con única respuesta.

Para administrar la prueba CBT lineal, basta con utilizar las funcionalidades por defecto de MOODLE en cuanto a selección de preguntas, número de ejercicios, porcentaje de peso en la calificación, escala de calificación y tiempo de duración de la prueba. Con la prueba CAT se utiliza un plugin o complemento gratuito desarrollado en lenguaje PHP por Middlebury College y Remote Learner[15], el cual permite definir el algoritmo de selección de preguntas, los criterios de parada, el nivel de dificultad para cada pregunta y visualizar los resultados de cada estudiante.

B. *Banco de preguntas*

Para la evaluación de diagnóstico se busca que las preguntas suministradas a los alumnos tengan un contexto similar al que se enfrentarán en la asignatura.

En el tema de conversión de unidades, se incluyen unidades de fuerza, longitud, presión, área, densidad, inercia y momento. Para el ejercicio de conversión de unidades de presión hidrostática se utiliza la ilustración de la Fig. 1. El alumno debe hacer consistente las unidades de densidad, gravedad y altura y luego realizar el producto de las tres cantidades para encontrar el valor de presión en las unidades requeridas. El concepto de presión hidrostática será estudiado en el curso para hallar las fuerzas generadas sobre la compuerta por la acción de un líquido.

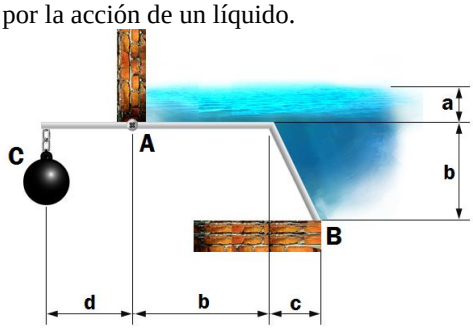

Fig. 1 Ejercicio unidades presión

En cuanto a geometría, se incluyen preguntas de área de figuras planas, volumen de sólidos regulares y semejanza entre triángulos, para ilustrar estas preguntas en la Fig. 2 se muestra la ilustración de un ejercicio de cerchas, donde por relación de triángulos el alumno debe encontrar la altura de la barra GC. Las cerchas se estudian en la asignatura como armaduras simples para analizar las fuerzas generadas en cada barra por las cargas externas.

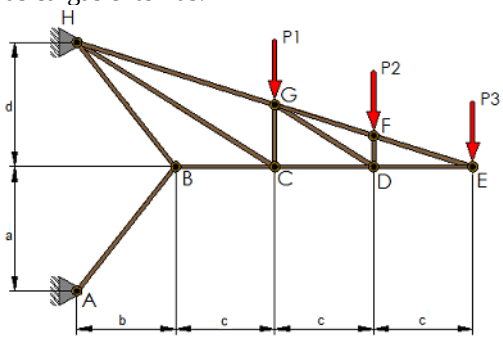

Fig. 2 Ejercicio semejanza entre triángulos

Con el fin de diagnosticar el manejo de relaciones entre los lados y ángulos de un triángulo se realizan preguntas de trigonometría, que incluyen la ley de Pitágoras, ángulos complementarios y suplementarios, suma de ángulos internos, funciones trigométricas, ley del seno y coseno. En la Fig. 3 se encuentra la ilustración de un modelo de ejercicio para aplicar la ley del coseno, la cota b, la cota a, el ángulo y la longitud de la cuerda BA son datos de entrada del ejercicio, el estudiante debe encontrar la longitud extendida del resorte AC; está configuración de cuerda y resorte se estudia en el tema de equilibrio estático.

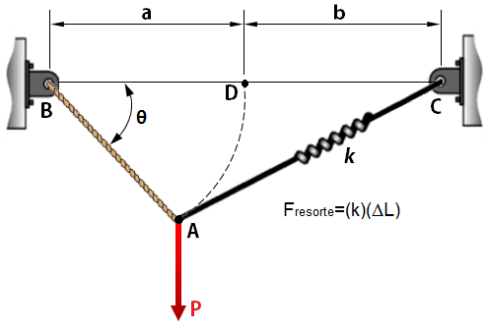

Fig. 3 Ejercicio trigonometría ley del coseno

En el diagnóstico de sistemas de ecuaciones se solicita al estudiante despejar una variable de expresiones obtenidas de un análisis estático, porque el estudiante al avanzar en el curso tendrá que plantear expresiones similares y determinar el valor de las variables. Se incluyen sistemas de una ecuación y una incógnita, y sistemas de dos ecuaciones simultáneas y dos incógnitas. En la Fig.4, se muestra la ilustración de un ejercicio cuyo objetivo es obtener el valor de la fuerza F, debido a que el valor de Fx, Fz y el ángulo son proporcionados como datos de entrada.

$$\left(\frac{Fx}{F}\right)^2 + \left(\frac{Fz}{F}\right)^2 + cos\beta^{\,2} = 1$$

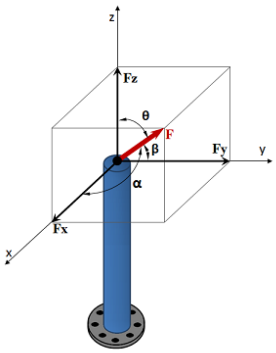

Fig. 4 Ejercicio una ecuación una incógnita

Luego del proceso de formulación y programación dinámica de las preguntas de diagnóstico se obtiene un banco de 52 preguntas distribuido en 17 preguntas de conversión de unidades, 11 de trigonometría, 14 de geometría y 10 preguntas de sistemas de ecuaciones.

C. Calibración de preguntas

Implementar y gestionar una prueba CAT requiere consideraciones cuidadosas y demandas psicométricas, el primer requerimiento es la validación de un banco de preguntas calibrado, porque los algoritmos para la selección de items requieren estimar la habilidad del estudiante cada que este responde una pregunta y la dificultad del item seleccionado debe estar en la misma escala [16].

La prueba de diagnóstico fue diseñada para el presente estudio por lo cual no se cuentan con registros previos de las respuestas a las preguntas, lo que permite aplicar IRT para determinar los parametros característicos de dificultad y discriminación de las preguntas. En consecuencia, los investigadores realizan la calibración inicial por medio del criterio de los profesores y el coordinador de la asignatura.

Para calibrar los niveles de dificultad de los ítems se tiene en cuenta el número de operaciones mínimo que el estudiante realiza en el ejercicio, la interpretación gráfica, el uso de tablas o fórmulas y la consistencia dimensional.

D. Algoritmo de selección de preguntas

La prueba CAT utiliza un algoritmo de selección iterativo según el desempeño del estudiante. Para el presente estudio se utiliza el algoritmo sugerido por B.D Wright en 1988 [17] y discutido por J. Linacre en el año 2000 [18].

A continuación se muestra el algoritmo utilizado el cual se basa en el modelo de Rasch:

1. Requerir próximo candidato: Asignar D=0, L=0, H=0, y R=0.
2. Encontrar próxima pregunta con la dificultad cercana (D).
3. Asignar D la actual calibración de esa pregunta.
4. Administrar esa pregunta.
5. Obtener respuesta.
6. Calificar esa respuesta.
7. Contar las preguntas tomadas: L = L + 1
8. Añadir la dificultad usada: H = H + D
9. Si la respuesta no es correcta, elegir dificultad pregunta siguiente: D = D - 2/L
10. Si la respuesta es correcta, contar respuestas correctas: R = R + 1
11. Si la respuesta es correcta, elegir dificultad pregunta siguiente: D = D + 2/L
12. Si no esta listo decisión de pasar/fallar, ir al paso 2.
13. Si esta listo decisión de pasar/fallar, calcular preguntas incorrectas: W = L - R
14. Estimar medida: B = H/L + log(R/W)
15. Estimar error estándar: S = sqrt[L/(R*W)]
16. Comparar medida B con T estándar pasar/fallar.
17. If (T - S) < B < (T + S), ir al paso 1.
18. If (B - S) > T, entonces pasa.
19. If (B + S) < T, entones falla.
20. Ir al paso 1.

E. Diseño de pruebas

La prueba CAT utiliza el banco completo de 52 preguntas y se administró antes de la prueba CBT, además no se encuentra limitada por tiempo. Los requerimientos técnicos de la prueba se definieron asi:

I. Nivel de dificultad inicial: 15
II. Nivel de dificultad más bajo: 1
III. Nivel de dificultad más alto: 52
IV. Mínimo número de preguntas respondidas: 5
V. Máximo número de preguntas respondidas: 20
VI. Error estándar de parada: 20%

Los criterios de parada del programa son el máximo número de preguntas y el error estándar, al cumplirse cualquiera de los dos criterios, el sistema aborta la iteración y muestra el resultado de la habilidad al estudiante en una escala de 0 a 5.

La prueba CBT cuya administración es secuencial y preestablecida por los coordinadores del estudio se activa al estudiante luego que responde la prueba CAT. Consta en total de 8 preguntas incluyendo 2 items de cada uno de los 4 temas y limitada en un tiempo de 30 minutos. Cada pregunta tiene el mismo valor porcentual en la calificación global de la prueba CBT.

Se extraen 8 preguntas del banco de 52 preguntas, los niveles elegidos son los siguientes:

Conversión de unidades: Nivel 5 y nivel 20
Trigonometría: Nivel 40 y nivel 41
Geometría: Nivel 22 y nivel 38
Solución de ecuaciones: Nivel 44 y nivel 49

F. Análisis de resultados

Para el análisis estadistico se recolectan los resultados de las dos pruebas para realizar las respectivas comparaciones; en primer lugar, se grafican los resultados de todos los items en una escala de 0 a 5, para cada prueba, para verificar las líneas de tendencia y la correlación. Para el estudio se busca definir si las dos pruebas miden la habilidad del estudiante en valores cercanos.

Se obtienen los valores promedio de la calificación global de cada prueba en una escala de 0 a 5 que concuerda con la escala utilizada en la universidad, de igual forma se extrae el tiempo promedio empleado para responder la prueba CBT y la prueba CAT de todo la población, y el promedio de preguntas respondidas por todos los estudiantes para cada caso.

III. Resultados

La Fig.5 muestra las curvas de comportamiento para ambas pruebas, se puede observar gráficamente una alta correlación entre los valores para los estudiantes con la habilidad más alta. El coeficiente de correlación global es de 0,605 que indica una correlación positiva moderada.

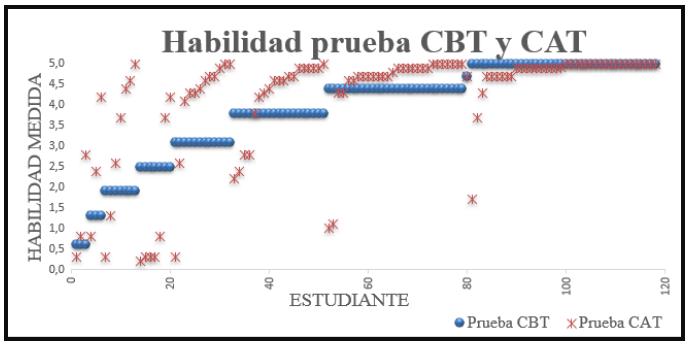

Fig. 5 Habilidad en prueba CBT y prueba CAT

La Tabla I contiene los valores promedios de habilidad medida en la escala de calificación de 0 a 5, el tiempo requerido para responder cada prueba y el número de preguntas respondidas en ambos casos.

TABLA I
Resultados promedio pruebas

	Prueba CAT	**Prueba CBT**
Nota promedio	4,11	3,93
Tiempo requerido (min)	44,04	23,19
No. Preguntas respondidas	7,48	8

Para el caso de la prueba CAT el máximo número de preguntas que respondió un estudiante fue 11 y el mínimo 6.

Por último, se muestra gráficos como ejemplo de la forma de presentación de los datos y el seguimiento que se realiza a la prueba CAT, que incluye la habilidad medida, el nivel elegido, nivel actual de dificultad y la franja del error estandar. En la Fig.6,7,8 y 9 se encuentra el comportamiento de la prueba CAT para cuatro estudiantes con distintos niveles de habilidad medidos. El mismo reporte se obtiene para toda la población, pero no se incluye en el presente artículo.

Así, en la Fig.7, al estudiante se le presentó al principio un ítem de dificultad 15 (la habilidad medida es 15 con un error estándar del 38,1%, que es la franja sombreada). La respuesta inicial fue incorrecta por lo que el algoritmo le presenta un ítem de dificultad 4 (habilidad estimada 7,8 error estándar 30,4%). La respuesta al segundo ítem fue correcta y el algoritmo le administró un ítem de dificultad 9 (habilidad estimada 13.57 con un error estándar de 27,3%). La prueba continua hasta la pregunta 7 que el algoritmo le suministra con un nivel de dificultad 23, estimando una habilidad de 25,52 con un error estándar del 19,8%.

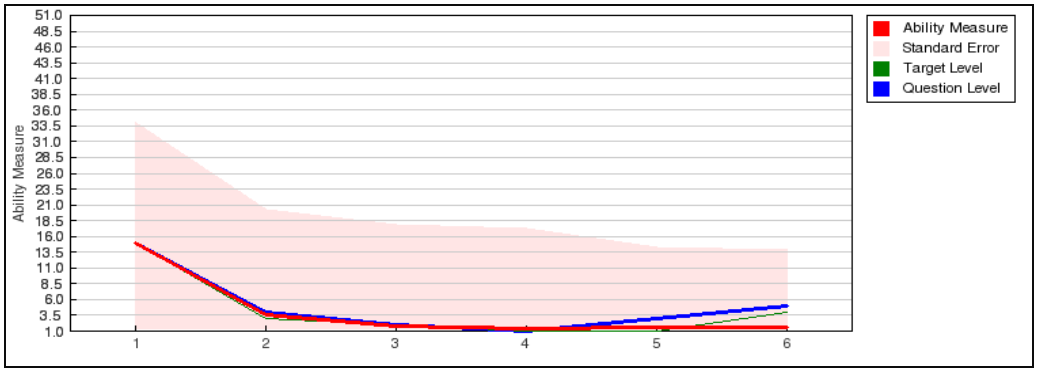

Fig. 6 Prueba CAT bajo desempeño 6 preguntas

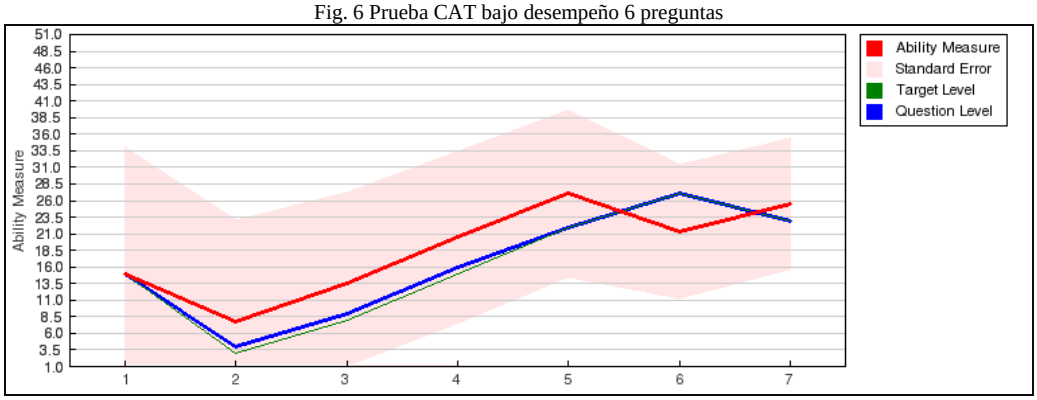

Fig. 7 Prueba CAT medio desempeño 7 preguntas

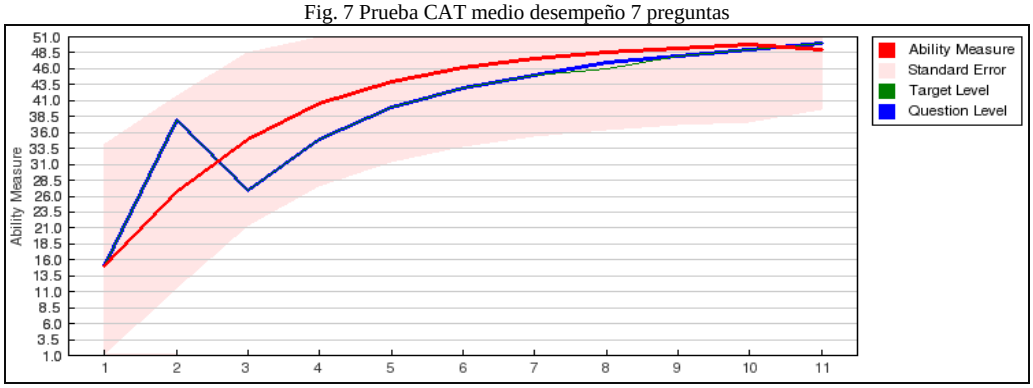

Fig. 8 Prueba CAT alto desempeño con fallo 11 preguntas

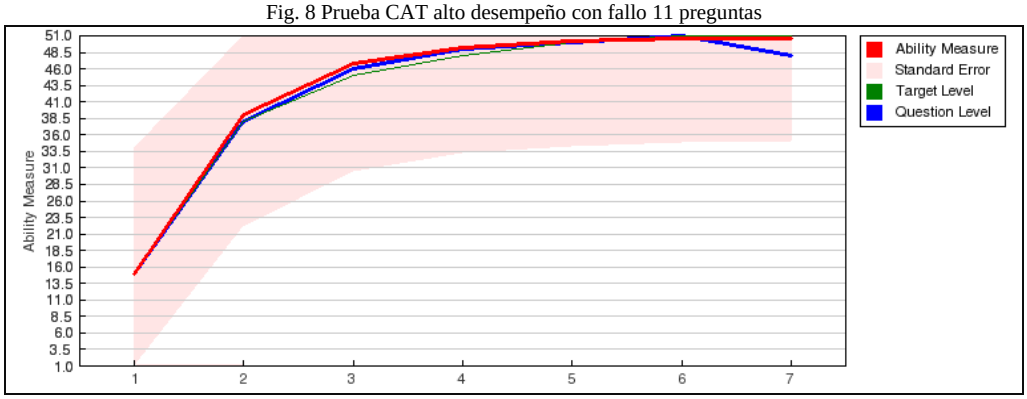

Fig. 9 Prueba CAT alto desempeño 7 preguntas

IV. Discusión

De la Fig.5 se obtienen las curvas de la habilidad medida para cada estudiante en ambas pruebas, Se puede observar una correlación positiva entre los resultados, esto es, que valores altos de la prueba CBT se correspondes con valores altos de la prueba CAT, principalmente para estudiantes con habilidades por encima de 4,0 (66 estudiantes) con una dispersión mayor para los estudiantes con habilidades menores. Además, la nota promedio de la Tabla I comprueba lo anterior, ya que la calificación global solo presenta 0,18 de diferencia entre la prueba CBT y la prueba CAT.

Con respecto al promedio de preguntas respondidas, en la Tabla I se muestra que en la prueba CAT los estudiantes respondieron menos preguntas que en la prueba CBT, lo que concuerda con uno de los aspectos planteados en la hipótesis, se verifica que la prueba adaptativa permite medir la habilidad del estudiante con menos preguntas, en este diseño de pruebas la diferencia fue del 0,52.

En la hipótesis del estudio se planteó que la prueba CAT mide la habilidad del estudio en menor tiempo que una prueba lineal, no obstante según los resultados de la Tabla I el promedio de tiempo requerido para la prueba CAT fue de 44,04 min mientras que en la segunda prueba de 23,19 min.

Esta diferencia de 20,85 min contradice la hipótesis en este aspecto. Para futuros estudios se busca revisar el motivo de este comportamiento que puede deberse a que la prueba CAT se administró primero que la prueba CBT y que la segunda estaba limitada a 30 min para todos los estudiantes.

El tercer aspecto de la hipótesis se refiere a la posibilidad de exigir al estudiante por medio de una dificultad creciente en la prueba CAT. Como se observa en la Fig.8 por ejemplo, el estudiante falló en una pregunta de nivel 38, pero el algoritmo permite que siga en un nivel inferior de dificultad 26 porque con anterioridad superó el nivel inicial de 15, el estudiante continua la prueba y con respuestas positivas sube hasta el nivel de dificultad de 50, cercano al máximo de la prueba. Se comprueba la hipótesis inicial en este aspecto.

Una ventaja de las pruebas adaptativas consiste en que se pueden almacenar los registros del estudiante para estabilizarlo en una habilidad determinada y aunque tenga un fallo en una pregunta puede seguir intentando de acuerdo a su historial y mejorar su calificación.

La prueba CAT se realizó con un error estándar de 20% debido a la longitud del banco de preguntas, futuros trabajos consisten en aumentar el número de preguntas para disminuir el error hasta niveles cercanos al 10%, aunque entre menor el error estándar requerido el alumno debe responder más preguntas.

El banco de preguntas inicial puede ser calibrado mediante la teoría de IRT porque ya se cuenta con respuestas a los ejercicios; dicha calibración permite definir la dificultad de la pregunta y su discriminación para establecer si la pregunta es pertinente al medir la habilidad del estudiante, en muchos casos la aplicación de IRT permitirá además descartar preguntas del banco.

Las pruebas adaptativas computarizadas se ajustan al desempeño del estudiante y miden su habilidad para responder un rango de dificultad. Para la enseñanza en ingeniería puede significar la realización de actividades evaluativas variadas y eficientes en cuanto tiempo de elaboración, calificación y exposición al usuario.

Trabajo futuros consisten en aplicar pruebas CAT en temas estudiados en Estática y comparar los resultados con registros de pruebas CBT de la asignatura elaborados desde el año 2011. Los resultados de este estudio permiten mejorar el diseño de pruebas CAT y CBT.

Reconocimientos

Los investigadores agradecen a la Universidad EAFIT de Medellín, Colombia por su apoyo logístico y financiero para llevar a cabo el presente estudio.

Referencias

[1] J. L. Restrepo Ochoa, “GENERADOR AUTOMÁTICO DE TAREAS COMO APOYO A LOS PROCESOS DE EVALUACIÓN, ASIGNATURA ESTÁTICA,” de Reunión Nacional ACOFI 2012, Medellín, 2012.

[2] J. L. Restrepo Ochoa, J. L. Barbosa Pérez y L. F. Zapata Rivera, “Resultados experimentales de la aplicación de un sistema de evaluación dinámico en la asignatura de Estática,” Latin American and Caribbean, vol. 7, nº 1, pp. 1-11, 2013.

[3] J. J. Tangarife Vélez, J. L. Barbosa Pérez y J. L. Restrepo Ochoa, “Efecto de un sistema interactivo con generación automática de ejercicios sobre el desempeño y la deserción de los estudiantes del curso de estática en ingeniería.,” Latin American and Caribbean, Santo Domingo, 2015.

[4] J. L. Restrepo Ochoa, J. L. Barbosa Pérez y A. Restrepo Cadavid, “MEDICIÓN DE LOS PARÁMETROS IRT DE UNA TAREA DINÁMICA EN LA ASIGNATURA ESTÁTICA,” de WEEF: Innovación en investigación y educación en ingeniería: Factores claves para la competitividad global, Cartagena, 2013.

[5] A. Restrepo Cadavid, J. L. Barbosa Pérez y J. L. Restrepo Ochoa, “Items’ difficulty level determination based on a Statics test with parameters,” de IEEE Frontiers in Education Conference, Madrid, 2014.

[6] A. Restrepo Cadavid, J. L. Barbosa Pérez y J. L. Restrepo Ochoa, “EVALUACIÓN DE LA DIFICULTAD DE PRUEBAS EN LA ASIGNATURA ESTÁTICA CON DIFERENTES TIPOS DE PREGUNTAS,” de EIEI 2014: Nuevos escenarios en la enseñanza de la ingeniería, Cartagena, 2014.

[7] L. Johnson, S. Becker y A. F. Victoria Estrada, “Horizon Report: 2014 Higher Education,” New Media Consortium, Austin, 2014.

[8] A. Dollár y P. Steif, “An Interactive, Cognitively Informed, Web-Based Statics Course,” International Journal of Engineering Education, vol. 24, nº 6, pp. 1129-1241, 2008.

[9] K. Gramoll, “eCourses: Online Engineering Course Management System,” University of Oklahoma, [En línea]. Available: https://www.ecourses.ou.edu/cgi-bin/navigation.cgi?course=st. [Último acceso: 25 Febrero 2016].

[10] A. Economides y C. Roupas, “Evaluation of Computer Adaptive Testing Systems,” International Journal of WebBased Learning and Teaching Technologies, vol. 2, nº 1, pp. 70-87, 2007.

[11] R. Conejo, B. Barros, E. Guzmán y J. Gálvez, “Formative evaluation of

the SIETTE collaborative testing environment," de Proceedings of the 16th International Conference on Computer in Education, ICCE08, Taipei, 2008.

[12] R. Conejo, E. Guzmán y M. Trella, "The SIETTE Automatic Assessment Environment," International Journal of Artificial Intelligence in Education, vol. 26, nº 1, pp. 270-292, 2016.

[13] D. J. Weiss, "Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education," Measurement and Evaluation in Counseling and Development, vol. 37, nº 2, pp. 70-84, 2004.

[14] W. Van der Linden y C. Glas, Elements of adaptive testing, Monterey: Springer, 2010.

[15] Middlebury College; Remote Learner, "Mod Adaptive Quiz," https://github.com/middlebury/moodle-mod_adaptivequiz., Middlebury, 2013.

[16] J. Barnard, "Implementing a CAT: The AMC experience," Journal of Computarized Adaptive Testing, vol. 3, nº 1, pp. 1-12, 2015.

[17] B. Wright, "Practical adaptive testing," Rasch Measurement Transactions, vol. 2, nº 2, p. 24, 1988.

[18] J. Linacre, "Computer-Adaptive Testing: A methodology wose time has come," MESA Memorandum No. 69, Seoul, 2000.

Aplicación de una prueba adaptativa computarizada en la asignatura Estática

Jorge Luis Restrepo Ochoa, Ph.D[1], Jaime Leonardo Barbosa Pérez, M.Sc[2], and Julian Arenas Berrio, Ing.[3]
[1,2,3]Universidad EAFIT, Colombia, jrestrep@eafit.edu.co, jbarbosa@eafit.edu.co, jarenas5@eafit.edu.co

***Abstract*– In the virtual platform for Statics course of the Engineering School at EAFIT University a Computer Based Test (CBT) and a Computerized Adaptive Test (CAT) were applied to measure the student's ability in previous subjects required for the Statics course. It was proved that the CAT's tests determine the student's ability with less amount of questions and take the examinee to the threshold of his capacity trough increasing levels of difficulty in contradiction to CBT's tests in which a fixed set of items is used. But in CAT's the duration time for the execution is higher. This paper presents the analysis of the collected data in both tests and it exposes the conclusions on the CAT's tests for the education in engineering.**

***Keywords*— Evaluation of Learning, Adaptability, Adaptive Tests, Computerized Tests, b-learning.**

***Resumen*– En la plataforma virtual para la asignatura Estática de la Escuela de Ingeniería de la Universidad EAFIT se aplicó una prueba basada en computador (CBT) y una prueba adaptativa computarizada (CAT) para medir la habilidad de los estudiantes en temas requeridos para el curso. Se comprobó que las pruebas CAT determinan la habilidad del estudiante en menor número de preguntas y llevan al examinado hasta el umbral de su capacidad a través de items con niveles de dificultad variables de acuerdo con el desempeño de cada estudiante, a diferencia de las pruebas CBT en las que se utiliza un conjunto de items fijo para todos los estudiantes. Sin embargo en las pruebas CAT el tiempo de aplicación fue mayor. Este trabajo presenta el análisis de los datos recolectados en ambas pruebas y expone las conclusiones sobre la prueba CAT para enseñanza en Estática.**

***Palabras clave*—Evaluación del aprendizaje, Adaptabilidad, Pruebas adaptativas, Pruebas computarizadas, b-learning**

I. Introducción

Debido al bajo rendimiento en la asignatura Estática para ingeniería en la universidad EAFIT, cuyo índice de reprobación de estudiantes matriculados era de 48.8% y la deserción se encontraba alrededor del 30% en los registros oficiales hasta 2011, la Escuela de Ingeniería implementó una plataforma virtual para mejorar los procesos de aprendizaje de los estudiantes y, por consiguiente su rendimiento académico [1] [2] [3].

La investigación alrededor de la plataforma virtual ha estado direccionada a la determinación de los parámetros característicos de dificultad y discriminación de las preguntas, y a su vez, la fiabilidad de los entrenamientos y pruebas suministradas a los estudiantes mediante la teoría de respuesta al ítem o IRT (Item Response Theory), con el objetivo de mejorar la calidad de los contenidos y del sistema [4] [5] [6].

Ahora bien, según la investigación sobre tendencias en la educación mundial, NMC HorizonReport: 2014 Higher Education Edition [7], para los próximos años se espera que la educación se adapte a las expectativas de los estudiantes y sus necesidades, tanto en el proceso de aprendizaje como en la evaluación de conocimientos. Actualmente en la asignatura Estática se aplican pruebas basadas en computador o CBT (Computerized based tests), las cuales se presentan de forma lineal y preestablecida al estudiante, pero con el fin de ajustar el proceso de evaluación con las tendencias mencionadas,en el presente estudio se aplica una prueba adaptativa computarizada o CAT (Computerized Adaptive Test) cuyas preguntas varían de acuerdo al desempeño de cada estudiante.

En la literatura se encuentran casos de implementación de plataformas virtuales para la enseñanza de Estática en ingeniería [8][9], pero no se encuentran registros con evaluaciones adaptativas para dicha asignatura. Las pruebas CAT, son utilizadas generalmente en contextos clínicos, exámenes estandarizados (e.g. GMAT, GRE, MCSE, TOEFL) y evaluación del desempeño en idiomas [10]. Para la evaluación del aprendizaje en ciencias básicas mediante pruebas tipo CAT, un punto de referencia es el sistema SIETTE [11] [12], desarrollado en la universidad de Málaga y extendido a otras universidades en España, no obstante no se muestra un estudio específico en asignaturas básicas en ingeniería como es Estática, el cual se aborda en esta investigación.

Al inicio del semestre 2016-1, se aplicó a 118 estudiantes de ingeniería civil, mecánica y de producción que cursan la asignatura Estática, dos pruebas para medir sus conocimientos en conversión de unidades, geometría, trigonometría y solución de ecuaciones, temas que los docentes identificaron como requeridos para lograr un buen desempeño en la asignatura. La primera prueba es adaptativa cuyos requisitos de aplicación Weiss [13] los enuncia de la siguiente forma: Banco precalibrado de preguntas, nivel de dificultad de la pregunta inicial, algoritmo iterativo de selección de la siguiente pregunta basado en la respuestas anteriores y un criterio de parada o terminación. La segunda prueba es de tipo lineal con ítems fijos aunque se varan los valores para cada estudiante con el fin de incrementar la seguridad de la prueba.

El objeto del estudio es comprobar que las pruebas CAT determinan la habilidad del estudiante con menor número de preguntas, en menor tiempo y con niveles de dificultad que llevan al examinado hasta el umbral de su capacidad tal como lo afirman diversos autores [10] [13] [14].

Digital Object Identifier (DOI): http://dx.doi.org/10.18687/LACCEI2016.1.1.280
ISBN: 978-0-9822896-9-3
ISSN: 2414-6390

II. Materiales y método

A. *Arquitectura del sistema*

La plataforma virtual se encuentra implementada en el sistema de gestión de contenidos de aprendizaje MOODLE, con licencia libre GNU (General public license) y albergada en los servicios de computación en la nube de AWS Amazon.

Para la generación de preguntas en formato compatible con MOODLE, se cuenta con una aplicación de escritorio desarrollada en lenguaje JAVA, cuyo propósito es programar el algoritmo de solución para cada pregunta y variar en un rango definido los datos de las variables relacionadas en el ejercicio. Mediante un estudio anterior se comprobó que variar los datos no modifica la dificultad de la pregunta [5], y en consecuencia, no cambia sustancialmente la habilidad a evaluar. Las preguntas utilizadas tanto en pruebas CBT como en pruebas CAT son de opción múltiple con única respuesta.

Para administrar la prueba CBT lineal, basta con utilizar las funcionalidades por defecto de MOODLE en cuanto a selección de preguntas, número de ejercicios, porcentaje de peso en la calificación, escala de calificación y tiempo de duración de la prueba. Con la prueba CAT se utiliza un plugin o complemento gratuito desarrollado en lenguaje PHP por Middlebury College y Remote Learner[15], el cual permite definir el algoritmo de selección de preguntas, los criterios de parada, el nivel de dificultad para cada pregunta y visualizar los resultados de cada estudiante.

B. *Banco de preguntas*

Para la evaluación de diagnóstico se busca que las preguntas suministradas a los alumnos tengan un contexto similar al que se enfrentarán en la asignatura.

En el tema de conversión de unidades, se incluyen unidades de fuerza, longitud, presión, área, densidad, inercia y momento. Para el ejercicio de conversión de unidades de presión hidrostática se utiliza la ilustración de la Fig. 1. El alumno debe hacer consistente las unidades de densidad, gravedad y altura y luego realizar el producto de las tres cantidades para encontrar el valor de presión en las unidades requeridas. El concepto de presión hidrostática será estudiado en el curso para hallar las fuerzas generadas sobre la compuerta por la acción de un líquido.

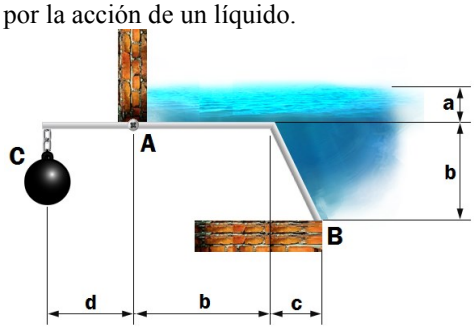

Fig. 1 Ejercicio unidades presión

En cuanto a geometría, se incluyen preguntas de área de figuras planas, volumen de sólidos regulares y semejanza entre triángulos, para ilustrar estas preguntas en la Fig. 2 se muestra la ilustración de un ejercicio de cerchas, donde por relación de triángulos el alumno debe encontrar la altura de la barra GC. Las cerchas se estudian en la asignatura como armaduras simples para analizar las fuerzas generadas en cada barra por las cargas externas.

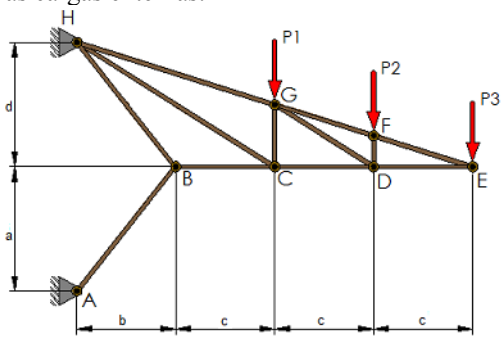

Fig. 2 Ejercicio semejanza entre triángulos

Con el fin de diagnosticar el manejo de relaciones entre los lados y ángulos de un triángulo se realizan preguntas de trigonometría, que incluyen la ley de Pitágoras, ángulos complementarios y suplementarios, suma de ángulos internos, funciones trigométricas, ley del seno y coseno. En la Fig. 3 se encuentra la ilustración de un modelo de ejercicio para aplicar la ley del coseno, la cota b, la cota a, el ángulo y la longitud de la cuerda BA son datos de entrada del ejercicio, el estudiante debe encontrar la longitud extendida del resorte AC; está configuración de cuerda y resorte se estudia en el tema de equilibrio estático.

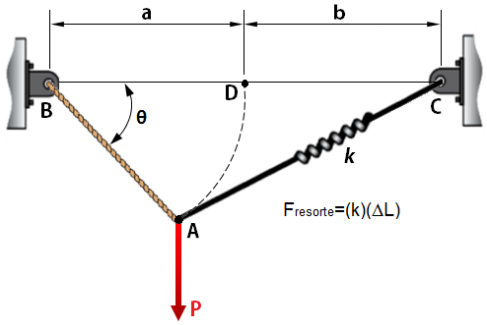

Fig. 3 Ejercicio trigonometría ley del coseno

En el diagnóstico de sistemas de ecuaciones se solicita al estudiante despejar una variable de expresiones obtenidas de un análisis estático, porque el estudiante al avanzar en el curso tendrá que plantear expresiones similares y determinar el valor de las variables. Se incluyen sistemas de una ecuación y una incógnita, y sistemas de dos ecuaciones simultáneas y dos incógnitas. En la Fig.4, se muestra la ilustración de un ejercicio cuyo objetivo es obtener el valor de la fuerza F, debido a que el valor de Fx, Fz y el ángulo son proporcionados como datos de entrada.

$$\left(\frac{Fx}{F}\right)^2 + \left(\frac{Fz}{F}\right)^2 + cos\beta^{\,2} = 1$$

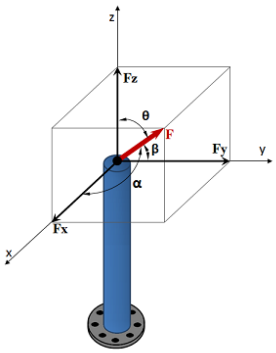

Fig. 4 Ejercicio una ecuación una incógnita

Luego del proceso de formulación y programación dinámica de las preguntas de diagnóstico se obtiene un banco de 52 preguntas distribuido en 17 preguntas de conversión de unidades, 11 de trigonometría, 14 de geometría y 10 preguntas de sistemas de ecuaciones.

C. Calibración de preguntas

Implementar y gestionar una prueba CAT requiere consideraciones cuidadosas y demandas psicométricas, el primer requerimiento es la validación de un banco de preguntas calibrado, porque los algoritmos para la selección de items requieren estimar la habilidad del estudiante cada que este responde una pregunta y la dificultad del item seleccionado debe estar en la misma escala [16].

La prueba de diagnóstico fue diseñada para el presente estudio por lo cual no se cuentan con registros previos de las respuestas a las preguntas, lo que permite aplicar IRT para determinar los parametros característicos de dificultad y discriminación de las preguntas. En consecuencia, los investigadores realizan la calibración inicial por medio del criterio de los profesores y el coordinador de la asignatura.

Para calibrar los niveles de dificultad de los ítems se tiene en cuenta el número de operaciones mínimo que el estudiante realiza en el ejercicio, la interpretación gráfica, el uso de tablas o fórmulas y la consistencia dimensional.

D. Algoritmo de selección de preguntas

La prueba CAT utiliza un algoritmo de selección iterativo según el desempeño del estudiante. Para el presente estudio se utiliza el algoritmo sugerido por B.D Wright en 1988 [17] y discutido por J. Linacre en el año 2000 [18].

A continuación se muestra el algoritmo utilizado el cual se basa en el modelo de Rasch:

1. Requerir próximo candidato: Asignar D=0, L=0, H=0, y R=0.
2. Encontrar próxima pregunta con la dificultad cercana (D).
3. Asignar D la actual calibración de esa pregunta.
4. Administrar esa pregunta.
5. Obtener respuesta.
6. Calificar esa respuesta.
7. Contar las preguntas tomadas: L = L + 1
8. Añadir la dificultad usada: H = H + D
9. Si la respuesta no es correcta, elegir dificultad pregunta siguiente: D = D - 2/L
10. Si la respuesta es correcta, contar respuestas correctas: R = R + 1
11. Si la respuesta es correcta, elegir dificultad pregunta siguiente: D = D + 2/L
12. Si no esta listo decisión de pasar/fallar, ir al paso 2.
13. Si esta listo decisión de pasar/fallar, calcular preguntas incorrectas: W = L - R
14. Estimar medida: B = H/L + log(R/W)
15. Estimar error estándar: S = sqrt[L/(R*W)]
16. Comparar medida B con T estándar pasar/fallar.
17. If (T - S) < B < (T + S), ir al paso 1.
18. If (B - S) > T, entonces pasa.
19. If (B + S) < T, entones falla.
20. Ir al paso 1.

E. Diseño de pruebas

La prueba CAT utiliza el banco completo de 52 preguntas y se administró antes de la prueba CBT, además no se encuentra limitada por tiempo. Los requerimientos técnicos de la prueba se definieron asi:

I. Nivel de dificultad inicial: 15
II. Nivel de dificultad más bajo: 1
III. Nivel de dificultad más alto: 52
IV. Mínimo número de preguntas respondidas: 5
V. Máximo número de preguntas respondidas: 20
VI. Error estándar de parada: 20%

Los criterios de parada del programa son el máximo número de preguntas y el error estándar, al cumplirse cualquiera de los dos criterios, el sistema aborta la iteración y muestra el resultado de la habilidad al estudiante en una escala de 0 a 5.

La prueba CBT cuya administración es secuencial y preestablecida por los coordinadores del estudio se activa al estudiante luego que responde la prueba CAT. Consta en total de 8 preguntas incluyendo 2 items de cada uno de los 4 temas y limitada en un tiempo de 30 minutos. Cada pregunta tiene el mismo valor porcentual en la calificación global de la prueba CBT.

Se extraen 8 preguntas del banco de 52 preguntas, los niveles elegidos son los siguientes:

Conversión de unidades: Nivel 5 y nivel 20
Trigonometría: Nivel 40 y nivel 41
Geometría: Nivel 22 y nivel 38
Solución de ecuaciones: Nivel 44 y nivel 49

F. Análisis de resultados

Para el análisis estadistico se recolectan los resultados de las dos pruebas para realizar las respectivas comparaciones; en primer lugar, se grafican los resultados de todos los items en una escala de 0 a 5, para cada prueba, para verificar las líneas de tendencia y la correlación. Para el estudio se busca definir si las dos pruebas miden la habilidad del estudiante en valores cercanos.

Se obtienen los valores promedio de la calificación global de cada prueba en una escala de 0 a 5 que concuerda con la escala utilizada en la universidad, de igual forma se extrae el tiempo promedio empleado para responder la prueba CBT y la prueba CAT de todo la población, y el promedio de preguntas respondidas por todos los estudiantes para cada caso.

III. Resultados

La Fig.5 muestra las curvas de comportamiento para ambas pruebas, se puede observar gráficamente una alta correlación entre los valores para los estudiantes con la habilidad más alta. El coeficiente de correlación global es de 0,605 que indica una correlación positiva moderada.

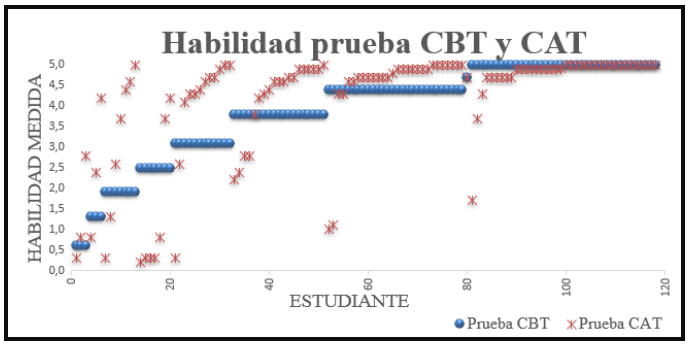

Fig. 5 Habilidad en prueba CBT y prueba CAT

La Tabla I contiene los valores promedios de habilidad medida en la escala de calificación de 0 a 5, el tiempo requerido para responder cada prueba y el número de preguntas respondidas en ambos casos.

TABLA I
Resultados promedio pruebas

	Prueba CAT	**Prueba CBT**
Nota promedio	4,11	3,93
Tiempo requerido (min)	44,04	23,19
No. Preguntas respondidas	7,48	8

Para el caso de la prueba CAT el máximo número de preguntas que respondió un estudiante fue 11 y el mínimo 6.

Por último, se muestra gráficos como ejemplo de la forma de presentación de los datos y el seguimiento que se realiza a la prueba CAT, que incluye la habilidad medida, el nivel elegido, nivel actual de dificultad y la franja del error estandar. En la Fig.6,7,8 y 9 se encuentra el comportamiento de la prueba CAT para cuatro estudiantes con distintos niveles de habilidad medidos. El mismo reporte se obtiene para toda la población, pero no se incluye en el presente artículo.

Así, en la Fig.7, al estudiante se le presentó al principio un ítem de dificultad 15 (la habilidad medida es 15 con un error estándar del 38,1%, que es la franja sombreada). La respuesta inicial fue incorrecta por lo que el algoritmo le presenta un ítem de dificultad 4 (habilidad estimada 7,8 error estándar 30,4%). La respuesta al segundo ítem fue correcta y el algoritmo le administró un ítem de dificultad 9 (habilidad estimada 13.57 con un error estándar de 27,3%). La prueba continua hasta la pregunta 7 que el algoritmo le suministra con un nivel de dificultad 23, estimando una habilidad de 25,52 con un error estándar del 19,8%.

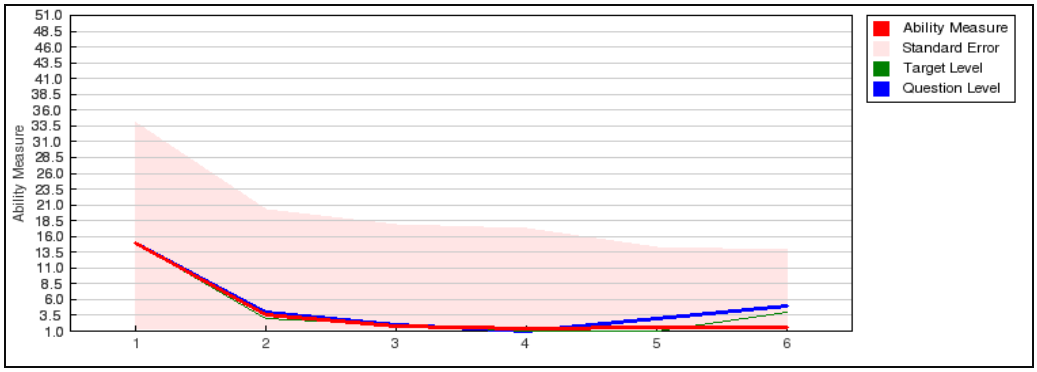

Fig. 6 Prueba CAT bajo desempeño 6 preguntas

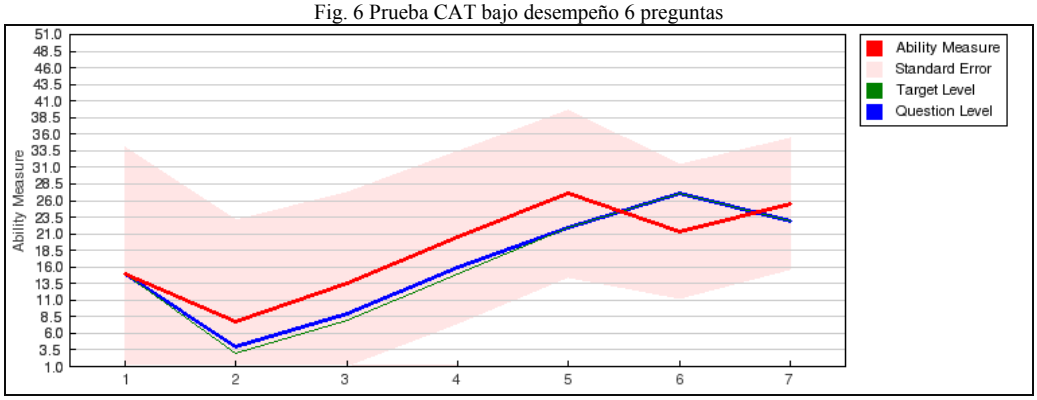

Fig. 7 Prueba CAT medio desempeño 7 preguntas

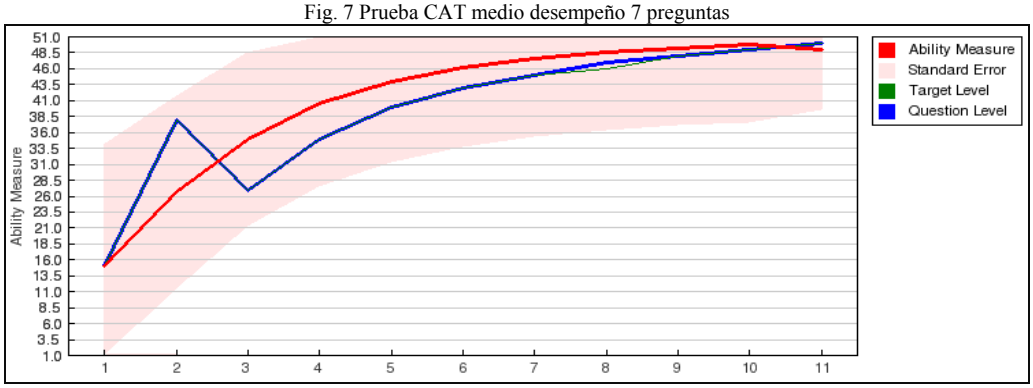

Fig. 8 Prueba CAT alto desempeño con fallo 11 preguntas

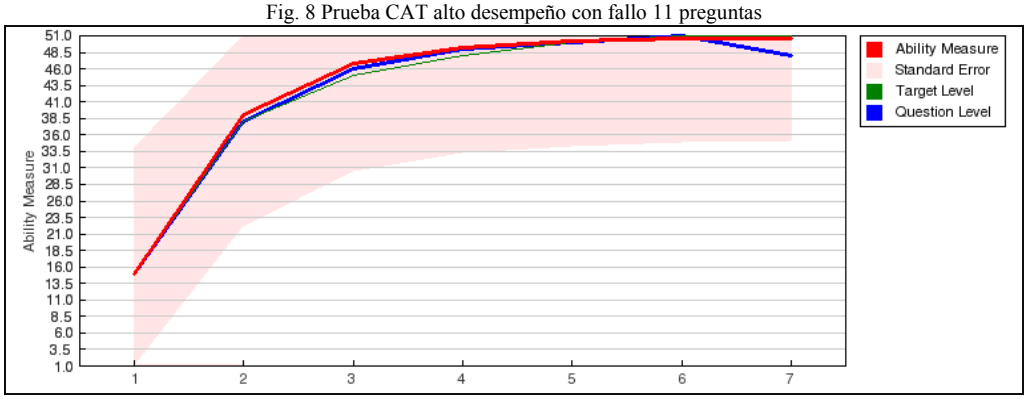

Fig. 9 Prueba CAT alto desempeño 7 preguntas

IV. Discusión

De la Fig.5 se obtienen las curvas de la habilidad medida para cada estudiante en ambas pruebas, Se puede observar una correlación positiva entre los resultados, esto es, que valores altos de la prueba CBT se correspondes con valores altos de la prueba CAT, principalmente para estudiantes con habilidades por encima de 4,0 (66 estudiantes) con una dispersión mayor para los estudiantes con habilidades menores. Además, la nota promedio de la Tabla I comprueba lo anterior, ya que la calificación global solo presenta 0,18 de diferencia entre la prueba CBT y la prueba CAT.

Con respecto al promedio de preguntas respondidas, en la Tabla I se muestra que en la prueba CAT los estudiantes respondieron menos preguntas que en la prueba CBT, lo que concuerda con uno de los aspectos planteados en la hipótesis, se verifica que la prueba adaptativa permite medir la habilidad del estudiante con menos preguntas, en este diseño de pruebas la diferencia fue del 0,52.

En la hipótesis del estudio se planteó que la prueba CAT mide la habilidad del estudio en menor tiempo que una prueba lineal, no obstante según los resultados de la Tabla I el promedio de tiempo requerido para la prueba CAT fue de 44,04 min mientras que en la segunda prueba de 23,19 min.

Esta diferencia de 20,85 min contradice la hipótesis en este aspecto. Para futuros estudios se busca revisar el motivo de este comportamiento que puede deberse a que la prueba CAT se administró primero que la prueba CBT y que la segunda estaba limitada a 30 min para todos los estudiantes.

El tercer aspecto de la hipótesis se refiere a la posibilidad de exigir al estudiante por medio de una dificultad creciente en la prueba CAT. Como se observa en la Fig.8 por ejemplo, el estudiante falló en una pregunta de nivel 38, pero el algoritmo permite que siga en un nivel inferior de dificultad 26 porque con anterioridad superó el nivel inicial de 15, el estudiante continua la prueba y con respuestas positivas sube hasta el nivel de dificultad de 50, cercano al máximo de la prueba. Se comprueba la hipótesis inicial en este aspecto.

Una ventaja de las pruebas adaptativas consiste en que se pueden almacenar los registros del estudiante para estabilizarlo en una habilidad determinada y aunque tenga un fallo en una pregunta puede seguir intentando de acuerdo a su historial y mejorar su calificación.

La prueba CAT se realizó con un error estándar de 20% debido a la longitud del banco de preguntas, futuros trabajos consisten en aumentar el número de preguntas para disminuir el error hasta niveles cercanos al 10%, aunque entre menor el error estándar requerido el alumno debe responder más preguntas.

El banco de preguntas inicial puede ser calibrado mediante la teoría de IRT porque ya se cuenta con respuestas a los ejercicios; dicha calibración permite definir la dificultad de la pregunta y su discriminación para establecer si la pregunta es pertinente al medir la habilidad del estudiante, en muchos casos la aplicación de IRT permitirá además descartar preguntas del banco.

Las pruebas adaptativas computarizadas se ajustan al desempeño del estudiante y miden su habilidad para responder un rango de dificultad. Para la enseñanza en ingeniería puede significar la realización de actividades evaluativas variadas y eficientes en cuanto tiempo de elaboración, calificación y exposición al usuario.

Trabajo futuros consisten en aplicar pruebas CAT en temas estudiados en Estática y comparar los resultados con registros de pruebas CBT de la asignatura elaborados desde el año 2011. Los resultados de este estudio permiten mejorar el diseño de pruebas CAT y CBT.

Reconocimientos

Los investigadores agradecen a la Universidad EAFIT de Medellín, Colombia por su apoyo logístico y financiero para llevar a cabo el presente estudio.

Referencias

[1] J. L. Restrepo Ochoa, "GENERADOR AUTOMÁTICO DE TAREAS COMO APOYO A LOS PROCESOS DE EVALUACIÓN, ASIGNATURA ESTÁTICA," de Reunión Nacional ACOFI 2012, Medellín, 2012.

[2] J. L. Restrepo Ochoa, J. L. Barbosa Pérez y L. F. Zapata Rivera, "Resultados experimentales de la aplicación de un sistema de evaluación dinámico en la asignatura de Estática," Latin American and Caribbean, vol. 7, nº 1, pp. 1-11, 2013.

[3] J. J. Tangarife Vélez, J. L. Barbosa Pérez y J. L. Restrepo Ochoa, "Efecto de un sistema interactivo con generación automática de ejercicios sobre el desempeño y la deserción de los estudiantes del curso de estática en ingeniería.," Latin American and Caribbean, Santo Domingo, 2015.

[4] J. L. Restrepo Ochoa, J. L. Barbosa Pérez y A. Restrepo Cadavid, "MEDICIÓN DE LOS PARÁMETROS IRT DE UNA TAREA DINÁMICA EN LA ASIGNATURA ESTÁTICA," de WEEF: Innovación en investigación y educación en ingeniería: Factores claves para la competitividad global, Cartagena, 2013.

[5] A. Restrepo Cadavid, J. L. Barbosa Pérez y J. L. Restrepo Ochoa, "Items' difficulty level determination based on a Statics test with parameters," de IEEE Frontiers in Education Conference, Madrid, 2014.

[6] A. Restrepo Cadavid, J. L. Barbosa Pérez y J. L. Restrepo Ochoa, "EVALUACIÓN DE LA DIFICULTAD DE PRUEBAS EN LA ASIGNATURA ESTÁTICA CON DIFERENTES TIPOS DE PREGUNTAS," de EIEI 2014: Nuevos escenarios en la enseñanza de la ingeniería, Cartagena, 2014.

[7] L. Johnson, S. Becker y A. F. Victoria Estrada, "Horizon Report: 2014 Higher Education," New Media Consortium, Austin, 2014.

[8] A. Dollár y P. Steif, "An Interactive, Cognitively Informed, Web-Based Statics Course," International Journal of Engineering Education, vol. 24, nº 6, pp. 1129-1241, 2008.

[9] K. Gramoll, "eCourses: Online Engineering Course Management System," University of Oklahoma, [En línea]. Available: https://www.ecourses.ou.edu/cgi-bin/navigation.cgi?course=st. [Último acceso: 25 Febrero 2016].

[10] A. Economides y C. Roupas, "Evaluation of Computer Adaptive Testing Systems," International Journal of WebBased Learning and Teaching Technologies, vol. 2, nº 1, pp. 70-87, 2007.

[11] R. Conejo, B. Barros, E. Guzmán y J. Gálvez, "Formative evaluation of

the SIETTE collaborative testing environment," de Proceedings of the 16th International Conference on Computer in Education, ICCE08, Taipei, 2008.

[12] R. Conejo, E. Guzmán y M. Trella, "The SIETTE Automatic Assessment Environment," International Journal of Artificial Intelligence in Education, vol. 26, nº 1, pp. 270-292, 2016.

[13] D. J. Weiss, "Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education," Measurement and Evaluation in Counseling and Development, vol. 37, nº 2, pp. 70-84, 2004.

[14] W. Van der Linden y C. Glas, Elements of adaptive testing, Monterey: Springer, 2010.

[15] Middlebury College; Remote Learner, "Mod Adaptive Quiz," https://github.com/middlebury/moodle-mod_adaptivequiz., Middlebury, 2013.

[16] J. Barnard, "Implementing a CAT: The AMC experience," Journal of Computarized Adaptive Testing, vol. 3, nº 1, pp. 1-12, 2015.

[17] B. Wright, "Practical adaptive testing," Rasch Measurement Transactions, vol. 2, nº 2, p. 24, 1988.

[18] J. Linacre, "Computer-Adaptive Testing: A methodology wose time has come," MESA Memorandum No. 69, Seoul, 2000.