La educación es de todos
Mineducación

SABER AL >DETALLE

EDICIÓN
06
Bogotá D.C.
Octubre de 2019
ISSN: 2590 - 4663
Publicación trimestral

Instituto Colombiano para la Evaluación de la Educación, ICFES
Oficinas: Calle 26 No. 69-76 . Torre 2, pisos 15 -18
Edificio Elemento, Bogotá . Colombia

Directora General: **María Figueroa Cahnspeyer**
Directora de Evaluación: **Natalia González Gómez**
Subdirectora de Análisis y Divulgación: **Ana María Restrepo Sáenz**

Subdirección de Estadísticas: **Jorge Mario Carrasco Ortíz**
Subdirección de Diseño de Instrumentos: **Javier Toro Baquero**
Coordinación General: **Dirección de Evaluación**

¿EN QUÉ CONSISTE LA APLICACIÓN DE PRE SABER 11° EN VERSIÓN ADAPTATIVA (CAT)?

icfes mejor saber

¿EN QUÉ CONSISTE LA APLICACIÓN DE PRE SABER 11º EN VERSIÓN ADAPTATIVA (CAT)?

El Icfes realizó un pilotaje del examen Pre Saber 11 bajo la metodología de prueba adaptativa por computador (CAT por sus siglas en inglés). Esta aplicación adaptativa emplea un número reducido de ítems o preguntas pertinentes para cada evaluado, los cuales permiten medir con mayor precisión su habilidad. Dado que la aplicación del examen Pre Saber 11 se realiza usualmente a través de lápiz y papel, fue necesario diseñar versiones de las pruebas para adecuarlas para su aplicación adaptativa. Así, el objetivo del piloto fue desarrollar las herramientas que soporten un examen adaptativo, de tal manera que se establezca una base para ser usada en exámenes posteriores. En esta edición profundizamos en la metodología que se expuso en el Saber al detalle sobre exámenes adaptativos por computador[1], y ahondamos en las particularidades que tuvo el desarrollo del pilotaje CAT para el examen PreSaber 11.

Para motivar la importancia del pilotaje del examen, analicemos las reacciones de los evaluados colombianos al presentar PreSaber 11 con versión en computador y con versión adaptativa en computador respecto a la presentación del examen en lápiz y papel. Se indagó sobre las ventajas y desventajas de estos tipos de aplicación frente a la metodología usual, y a través de un análisis cualitativo se obtuvo la nube de palabras de la figura 1. Se observa que en las ventajas se emplea una gran cantidad de palabras, en la que se resaltan "fácil", "mejor", "sencillo", "estudiante". Por su parte, en las desventajas se relacionan con aspectos relacionados con la herramienta como "pantalla", "brillo", "lectura" y aspectos por mejorar como "devolver", "subrayar", "letra" .

FIGURA 1. Análisis textual de ventajas y desventajas de la aplicación CAT o en versión computarizada frente a lápiz y papel

Fuente: Icfes, 2019

1. Para profundizar sobre CAT y sus generalidades refiérase a la Edición 2 del Boletín Saber al Detalle ¿En qué consiste la aplicación de pruebas adaptativas por computador (CAT) para las pruebas Saber?

¿Cuál es el examen PreSaber 11?

El examen Pre Saber 11 tiene como objetivo que los estudiantes de grado 10° o aquellos interesados se familiaricen con la estructura y las condiciones de aplicación del Examen de Estado Saber 11. Por lo cual, es un examen que le permite a los estudiantes poner a prueba las habilidades desarrolladas en la etapa escolar, e identificar sus fortalezas y los aspectos por mejorar. PreSaber 11 tiene la misma estructura de Saber 11, por lo cual está compuesto por las pruebas de lectura crítica, matemáticas, ciencias naturales, sociales y ciudadanas e inglés. Cada una de estas pruebas se diseña utilizando como referente nacional los Estándares Básicos de Competencias definidos por el Ministerio de Educación, y sus respectivas matrices de referencia se desarrollan bajo un Diseño Basado en Evidencias, en el que se observan las competencias evaluadas de cada una de las cinco pruebas y sus características particulares.

En la versión de lápiz y papel, cada evaluado presenta una forma de medición predefinida a ser desarrollada en dos sesiones, la cual se configura a partir de un diseño en bloques incompletos balanceados (BIBs)[2]. Recordemos que, bajo este diseño, se cuenta con diferentes formas de medición del examen y tal número es definido. La primera diferencia que encontramos con la aplicación en CAT es que la forma medición que presenta cada estudiante se podría considerar personalizada, en la medida en que los ítems se seleccionan de acuerdo con la habilidad observada de cada estudiante y no están predefinidos en agrupaciones de bloques, como lo es en el caso de lápiz y papel. Es importante señalar que para lograr comparabilidad entre lápiz y papel y CAT se requiere que lo que se pretende evaluar sea lo mismo, esto es garantizar las mismas especificaciones de las pruebas entre modalidades de aplicación. Más adelante observaremos que el número de ítems con un aplicativo CAT es menor por prueba respecto a la prueba a lápiz y papel.

A la hora de formular el diseño bajo el aplicativo de CAT es necesario contemplar las especificaciones del examen, que comprenden las características particulares de cada una de las cinco pruebas. Para el examen PreSaber 11 la mayoría de las pruebas agrupan ítems que comparten enunciados similares y que contienen situaciones utilizadas en las pruebas, las cuales se conocen como contextos. En el caso de la prueba Saber 11, estos contextos pueden ser familiares o personales, laborales u ocupacionales, comunitarios o sociales, matemáticos o científicos[3].

2. Para profundizar el diseño del armado, sus supuestos y las especificaciones de las pruebas Saber refiérase a la Edición 5 del Boletín Saber al Detalle ¿Qué diseño del armado se emplea en el Icfes para medir las pruebas Saber?

3. Si quiere leer más información sobre las particularidades de la prueba Saber 11, remítase a las guías de orientación en **https://www2.icfes.gov.co/web/guest/acerca-examen-saber-11#Informaci%C3%B3n%20general**

¿Por qué PreSaber 11° CAT?

Alrededor del mundo se han realizado desarrollos adaptativos de pruebas que solían ser aplicadas en lápiz y papel y toman en cuenta la existencia de mejoras tecnológicas y computaciones aplicadas a la estadística y a la psicometría. En esa línea, desde el Icfes se articularon estrategias innovadoras técnicas y tecnológicas en la aplicación electrónica de la prueba PreSaber 11 con el algoritmo de selección de ítems adaptativos, que suministre los ítems más pertinentes para cada evaluado y cuenta con mayor precisión en la estimación de la habilidad de los evaluados. Buscando que los estudiantes se acerquen al examen Saber 11 y que pongan a prueba sus habilidades desarrolladas en contextos educativos, se realizó el diseño de cinco motores adaptativo que fueran capaces de satisfacer tal necesidad, y que a la vez garantizaran que los ítems disponibles fueran los adecuados para ser aplicados con CAT. Más aún, cuando en el contexto nacional no se ha realizado estudios sobre el uso de pruebas con motor adaptativo. El método de calificación continúa siendo el empleado en la aplicación de lápiz y papel, teoría respuesta al ítem (TRI), el cual permite modelar el comportamiento de la habilidad de los evaluados utilizando información relacionada con las características de los ítem, como por ejemplo la dificultad, la discriminación y el azar.

¿Qué elementos son importantes en CAT para PreSaber 11°?

Para cumplir el objetivo del desarrollo de la aplicación CAT, se requiere analizar las especificaciones de cada prueba, puesto que estas forman parte de las características del diseño básico para su versión adaptativa. Tales especificaciones pueden ser las competencias, las afirmaciones o módulos, por ejemplo[4]. En el examen PreSaber 11 se realiza una selección adaptativa tanto a nivel de ítems como a nivel de agrupaciones de ítems, por lo cual resulta ser un diseño de CAT híbrido. Esta distinción sobre la selección adaptativa es clave para el examen, ya que la mayoría de las pruebas agrupan los ítems en contextos, y por tanto se debe asegurar dicha particularidad de la prueba. Tomando en cuenta estas características, el Icfes desarrolló un diseño adaptativo independiente para cada una de las 5 pruebas.

Una vez analizado el diseño de cada una de las pruebas, el funcionamiento óptimo de un examen bajo esta metodología está sujeto a: (1) el número de ítems disponibles para ser usados en el examen que logren medir diferentes niveles de habilidad en cada una de las pruebas, es decir el Banco de ítems; y (2) a la calidad psicométrica de los ítems. En este sentido, el Banco de ítems debe contar con ítems de diferentes dificultades, pues el motor adaptativo selecciona el ítem más adecuado para cada evaluado según su nivel de habilidad[5].

Como los ítems que son aplicados a un evaluado pueden ser aplicados a otros evaluados según su habilidad estimada, el banco de ítems resulta ser el lugar donde reposa el conjunto de los ítems requeridos para todos los

4. Para profundizar sobre el Diseño Centrado en Evidencias refiérase a **https://www.icfes.gov.co/documents/20143/516332/Guia+introductoria+al+dise%C3%B1o+centrado+en+evidencias+2018.pdf**

5. Recordemos que en TRI la habilidad y la dificultad se miden bajo la misma escala.

evaluados. Para definir el banco de ítems requerido para CAT fue necesario realizar un estudio de simulación con los parámetros de los ítems existentes, que se conocen como ítems precalibrados, y validarlos a la luz de las características de un banco de ítems ideal, teniendo en cuenta la función de información del examen ¿la recuerdan?[6]. Se realizaron análisis sobre la habilidad, la función de información y las distribuciones de información sobre rangos de dificultad. Típicamente, los ítems tienen un máximo de información alrededor de un rango de habilidades, por lo cual la escala de habilidad se definió por segmentos cuyo ancho es el rango medio cercano a la máxima información (Reckase, 1976).

¿Por qué se aplican menos ítems en CAT respecto a lápiz y papel?

La versión de PreSaber 11 de lápiz y papel logra medir con un tamaño de 221 ítems las habilidades de los evaluados. Con ese referente, se busca definir el tamaño del examen que garantice la medición de dichas habilidades con la versión adaptativa, para lo cual se debe garantizar que las características de la aplicación en lápiz y papel sean similares a su versión adaptativa. Una de esas características es la función de información. Brevemente, recordemos que en la publicación anterior, la función de información en modelos TRI es una herramienta estadística importante para calcular la precisión en las estimaciones, pues a mayor información es menor el error estándar de estimación. En ese sentido, se busca encontrar una función de información del examen versión CAT similar a la función correspondiente para la versión en lápiz y papel, lo cual nos indica unos errores de medición similares entre aplicaciones.

A través de un análisis de longitud del examen es posible analizar las funciones de información de los dos tipos de aplicación del examen, y definir el tamaño requerido para cada prueba de PreSaber 11 en CAT que permita garantizar una función de información del examen similar a la estimada en lápiz y papel. En la tabla 1 se observa el número de ítems por prueba en la versión de lápiz y papel que contempla las especificaciones de contenido por prueba, con un total de 221 ítems por examen. Según los análisis realizados, se estima que la longitud adecuada de ítems por prueba que se aproxima con precisión a la versión de lápiz y papel está entre 15 a 20 ítems. En esa línea, el número mínimo de ítems para cumplir con los requerimientos por competencia, componente o afirmación del examen PreSaber 11 puede variar entre 85 a 94 ítems, con lo cual se obtiene una precisión similar a la prueba de lápiz y papel. De ahí que el número de ítems en la aplicación de CAT sea menor que en lápiz y papel.

TABLA 1. Descripción de la longitud del examen Pre Saber 11° en aplicación papel y lápiz.

Prueba	Preguntas por prueba versión lápiz y papel	Preguntas por prueba versión CAT
Matemáticas	44	15
Lectura Crítica	36	15
Sociales y Ciudadanas	44	20
Ciencias Naturales	52	20
Inglés	45	15-23
Total	221	85-92

Fuente: Icfes, 2019

6. En la publicación anterior mencionamos que existe un rango de habilidad particular en el que hay máxima información del examen y a la vez menores errores estándar.

¿Cómo se realiza el diseño y el algoritmo adaptativo?

Como sabemos, bajo la metodología de CAT se diseña un algoritmo adaptativo en el que se incluyen ítems pre calibrados, un criterio de inicio, un criterio de elección de ítems, y un criterio de parada. Para la aplicación de PreSaber 11 CAT, se adaptó la plataforma PLEXI (Plataforma de Presentación de Exámenes del Icfes) y se diseñaron los algoritmos adaptativos con base en los siguientes criterios que se muestran a continuación.

1. Criterio de inicio

En el caso de Pre Saber 11, se seleccionó un criterio de inicio en el que se asume que la distribución de habilidades de cada evaluado es igual al promedio de la habilidad de los evaluados, que se define en cero con una desviación estándar de 1. Se inicia con una pregunta de dificultad media, y según la respuesta del evaluado se aplica el siguiente ítem informativo para el evaluado.

2. Criterio de elección de ítems

Una vez el evaluado dé respuesta al ítem de inicio y se haya estimado su habilidad a través de TRI, el algoritmo adaptativo suministra el siguiente ítem. Se busca que el ítem seleccionado sea lo más informativo posible, dado el rango de habilidad estimada en la que se encuentre el evaluado. Como los ítems están pre calibrados, se conoce el comportamiento psicométrico de los ítems, que en un modelo 3PL corresponde a la discriminación, la dificultad y el azar[7]. El ítem seleccionado por el algoritmo adaptativo resulta ser aquel que tiene la dificultad más cercana a la habilidad estimada del estudiante, ya que tanto la habilidad como la dificultad se encuentran en la misma escala. Según la prueba y sus características, es posible que la adaptación se realice a nivel ítem o conjunto de ítems, contemplando la respuesta obtenida por el estudiante en el ítem previamente suministrado. En este sentido, la habilidad estimada del evaluado se actualiza con cada ítem administrado de manera iterativa.

3. Criterio de parada

Como mencionamos en la sección anterior, se definió un rango del número de ítems por prueba para contemplar las características de cada una. Por lo cual el criterio de parada corresponde a un número definido de ítems administrados, que se conoce como longitud finita, que finaliza cuando se hayan completado el número de ítems señalados para cada una de las cinco pruebas.

¿Qué hallazgos se encontraron en el piloto PreSaber 11 CAT?

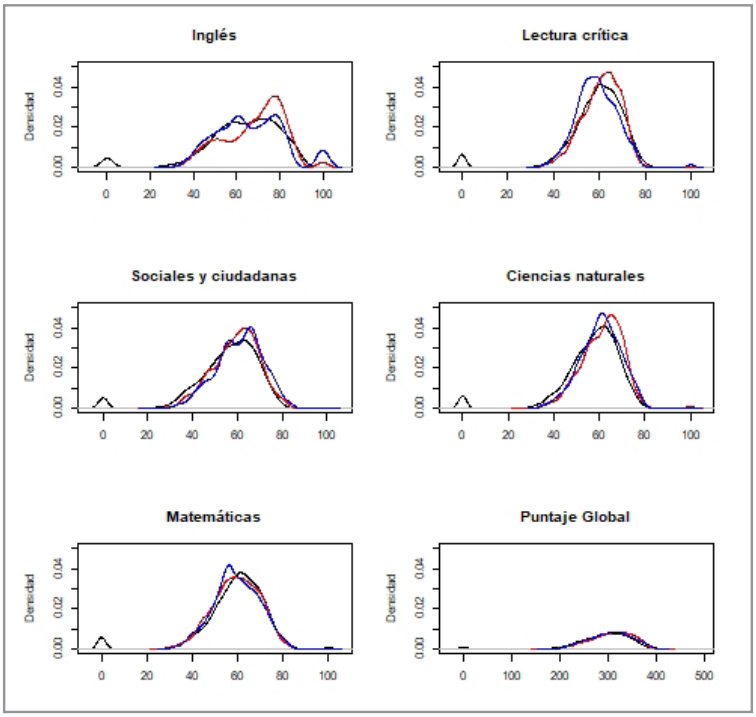
Teniendo en cuenta la existencia de versiones de aplicación distintas a papel y lápiz, surge la pregunta alrededor de la existencia de diferencias en el desempeño de los evaluados entre formas de aplicar el examen. En particular, alrededor de 3 modalidades de aplicación: papel y lápiz; la versión de papel y lápiz en computador, que llamamos electrónico; y la versión adaptativa computarizada, CAT. Los análisis de interés buscan comparan los resultados entre (1) lápiz y papel y CAT, y (2) entre lápiz y papel y electrónico. Para investigar sobre posibles diferencias, se realizó un muestreo para aplicar el examen PreSaber en modalidad electrónica y CAT a algunos estudiantes

7. Para profundizar sobre la metodología de calificación y sus supuestos refiérase a la Edición 1 del Boletín Saber al Detalle ¿Cómo se generan los puntajes en las pruebas Saber del Icfes?

en varios municipios de Colombia, y poder contrastar de manera representativa estos resultados con el desempeño obtenido en la aplicación usual del examen que es lápiz y papel. Así mismo, evaluar el impacto de la herramienta digital sobre el desempeño de los evaluados. Cabe señalar que se realizó esta aplicación a la misma población para poder realizar las comparaciones entre modalidades de aplicación.

A continuación, analizaremos descriptivamente si las habilidades de los estudiantes se comportan de manera similar entre las versiones del examen a lo largo de los diferentes niveles de habilidad, que es nuestra variable de interés. Para ello, nos remitimos a analizar las distribuciones para cada modalidad de aplicación (ver figura 2). De manera general observamos que las distribuciones de resultados son similares por prueba entre los distintos tipos de aplicación, lo cual nos indica que hay consistencia alrededor del puntaje promedio de los evaluados y la dispersión de los resultados, que es la desviación estándar, sin importar si la aplicación se realiza a través de papel y lápiz, electrónica o CAT.

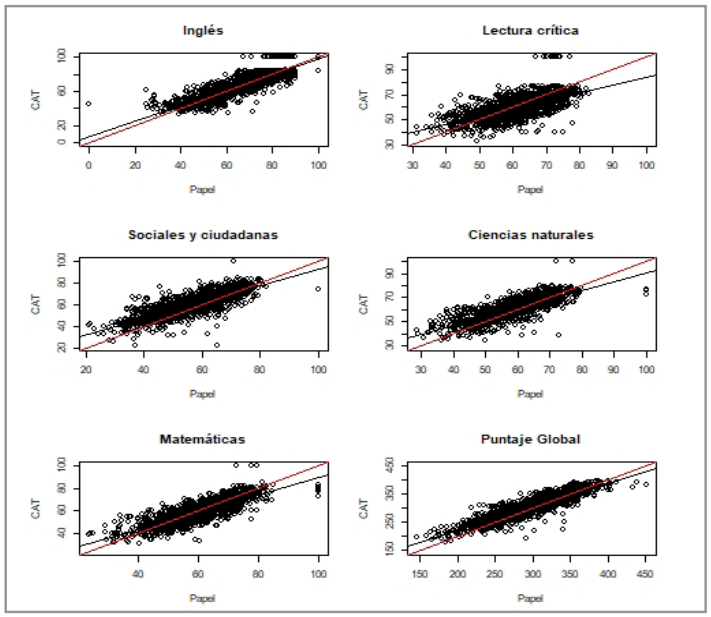
FIGURA 2. Distribución de puntajes por prueba y versión de aplicación del examen PreSaber 11.



Fuente: Icfes, 2019

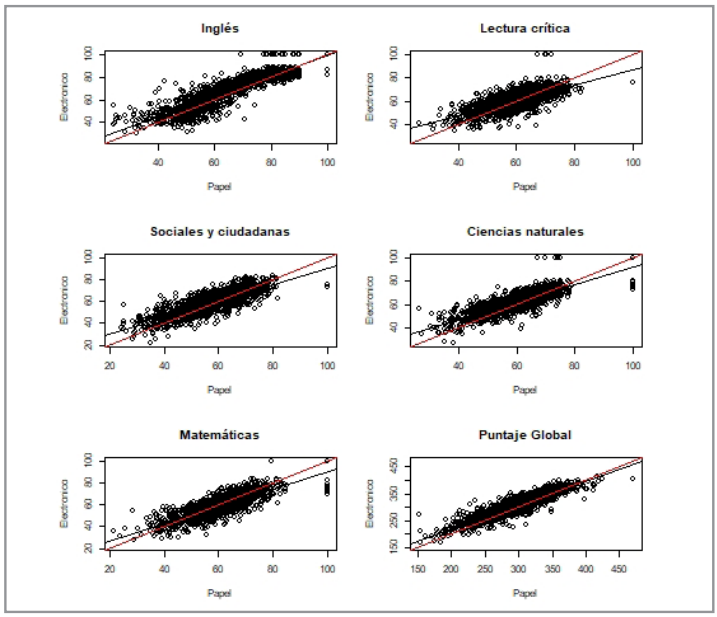
Luego, analicemos si los resultados de los evaluados son similares a los estimados a través de la aplicación en lápiz y papel sin importar el tipo de aplicación del examen computarizado. Si dicha hipótesis es cierta se esperaría que, en promedio, existiera una asociación positiva de los resultados al compararlos por tipo de aplicación, y de manera gráfica esto se evidenciaría con líneas de regresión con tendencia similar. En la figura 3 observamos la comparación de los resultados entre lápiz y papel y CAT, y podemos afirmar dicho fenómeno, ya que la asociación que hay entre los puntajes obtenido es similar. En general, se observa que dicha relación es positiva y fuerte, pues los datos se ajustan a la línea de estimación. Estos hallazgos son similares al comparar los resultados por prueba entre lápiz y papel y electrónico (figura 4).

FIGURA 3. Relación de puntajes por prueba entre aplicación lápiz y papel y CAT para la prueba PreSaber 11.



Fuente: Icfes, 2019

FIGURA 4. Relación de puntajes por prueba entre aplicación lápiz y papel y CAT para la prueba PreSaber 11.



Fuente: Icfes, 2019

Como las líneas de regresión por aplicación no se sobreponen, posiblemente existan diferencias de resultado por tipo aplicación del examen. Sin embargo, ¿serán estadísticamente grandes dichas diferencias? Al realizar estimaciones de las diferencias promedio entre aplicaciones por prueba para la misma población se encuentra que son pequeñas. Por un lado, las diferencias respecto a CAT oscilan entre 1 punto y 3,43 en valor absoluto, y respecto a electrónica entre 0,56 y 2,78 puntos en valor absoluto, y teniendo en cuenta que la desviación estándar por prueba son 10 puntos, las diferencias encontradas representan a lo sumo un tercio de la dispersión de los datos. Ahora bien, para cuantificar en contexto si esa diferencia para la misma población es grande, se calcula la estadística sobre el tamaño del efecto que tiene en cuenta tanto la media y la dispersión de los datos, y se concluye que en ambos casos la diferencia pequeña, pues el estadístico D de Cohen es menor a 0,5.

En conclusión, el piloto PreSaber 11 en versión adaptativa muestra que es eficiente para medir las habilidades de los evaluados y estadísticamente se observan resultados similares a los obtenidos en la versión usual de lápiz y papel aplicados a la misma población. La ventaja de aplicar el examen PreSaber 11 en CAT es la pertinencia para cada evaluado, pues se administran los ítems según su nivel de habilidad. Como se observa, son varias las ventajas que ofrecen las pruebas adaptativas, las cuales se fortalecen con el desarrollo de nuevas técnicas o herramientas tecnológicas, que permiten elaborar pruebas de mejor calidad.

Bibliografía

Icfes (2019) *Proyecto estratégico Prueba Adaptativa Computarizada – CAT.* Examen PreSaber 11.

Reckase, M. D. (1976). *The effect of item pool characteristics on the operation of a tailored testing procedure. Paper presented at the spring meeting of the Psychometric Society*, Murray Hill, NJ.

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Mislevy, R. J., Steinberg, L., & Thissen, D. (2014). *Computerized Adaptive Testing: a primer 7* by Howard Wainer with Neil J. Dorans. London and New York: Routledge Taylor and Francis Group.

SABER AL DETALLE

¿EN QUÉ CONSISTE LA APLICACIÓN DE PRE SABER 11º EN VERSIÓN ADAPTATIVA (CAT)?

@icfescol | ICFES | icfescol | YouTube: ICFES

Instituto Colombiano para la Evaluación de la Educación, ICFES
Oficinas: Calle 26 No. 69-76 . Torre 2, pisos 15 -18
Edificio Elemento, Bogotá . Colombia