La educación es de todos
Mineducación

SABER AL DETALLE

EDICIÓN
05
Bogotá D.C.
Julio 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 PRUEBAS ADAPTATIVAS POR COMPUTADOR (CAT) PARA LAS PRUEBAS SABER?

icfes mejor saber

¿EN QUÉ CONSISTE LA APLICACIÓN DE PRUEBAS ADAPTATIVAS POR COMPUTADOR (CAT) PARA LAS PRUEBAS SABER?

Para la medición con pruebas estandarizadas, se han usado instrumentos como papel y lápiz, y recientemente, exámenes electrónicos. A través de un medio tecnológico, no solo es posible hacer una prueba de manera digital equivalente a una en papel y lápiz, sino también, se pueden adoptar metodologías para hacer una prueba más eficiente, como el caso de pruebas adaptativas por computador (CAT por sus siglas en inglés), que permite el ajuste de la dificultad de las preguntas a las habilidades de los evaluados. Como se verá más adelante, este tipo de pruebas mejora la calidad de la medición, ya que usa algoritmos de administración eficiente de preguntas y aumenta la seguridad. Esta eficiencia está sujeta al tamaño del Banco de ítems y a la calidad psicométrica de los ítems, de tal manera que estos sean capaces de medir todo el rasgo latente de habilidad; esto es, que se cuente con ítems de calidad para medir tanto habilidades bajas como habilidades altas. Recientemente, el Icfes realizó el pilotaje del examen PreSaber bajo la metodología de prueba adaptativa, para lo cual desarrolló un diseño de cada uno de las 5 componentes del examen.

Durante una prueba adaptativa se administran los ítems según la respuesta del evaluado a cada uno. La adaptación requiere la estimación de habilidad en cada paso, a través de teoría respuesta al ítem (TRI), con dos propósitos: 1) actualizar continuamente la habilidad del evaluado, y 2) determinar la selección del siguiente ítem administrado. De esta manera, existen diferentes trayectorias de dificultad de los ítems administrados dentro de la prueba, que se ajustan a las competencias de cada estudiante a través de algoritmos computacionales.

1. ¿Qué es CAT?

Las pruebas adaptativas son una optimización de la versión computarizada equivalente a las pruebas de papel y lápiz, que consisten en seleccionar la siguiente pregunta de acuerdo a la habilidad actual del evaluado, iterativamente, durante la secuencia de preguntas del examen. Entre los diversos objetivos de toda prueba estandarizada está minimizar o reducir los errores de medición con el objetivo de tener una mejor estimación de la habilidad. Por lo tanto, en el caso de una prueba adaptativa, se trata de tener la mayor confiabilidad y conseguir la mayor cantidad de información posible después de administrar cada ítem, sin necesidad de administrar un examen de tamaño considerable.

En términos prácticos, un motor adaptativo replica de manera automática y eficiente lo que haría un evaluador experto cuando suministra preguntas a los evaluados, teniendo en cuenta su nivel de competencia o habilidad. Lo anterior implica seleccionar de manera iterativa los ítems pertinentes, es decir, aquellos que nos den la mayor y mejor información para así medir la habilidad del evaluado. Por ejemplo, si al suministrar una pregunta a un evaluado este la responde correctamente, el motor adaptativo seleccionará y suministrará un ítem de mayor dificultad, puesto que no brindaría información adicional sobre la habilidad del evaluado suministrar una pregunta similar en dificultad que la anteriormente suministrada.

Para que el algoritmo sea eficiente se debe considerar el tamaño del Banco de ítems, que es el conjunto de preguntas disponibles para un examen empleadas para medir la habilidad en distintos niveles. En ese sentido, es deseable contar con un conjunto de ítems que apunten a medir los distintos puntos del rasgo latente, pues el motor adaptativo selecciona y administra las preguntas de ese banco a los evaluados, condicional a las respuestas de ítems previamente suministrados. En términos de precisión, para estimar la habilidad de los evaluados se debe contar con preguntas de alta calidad, más aún cuando en CAT se usa una cantidad menor de ítems a la que se requiere con pruebas en papel y lápiz.

Como se observa, un ítem da mucha información en términos de sus características, así que se debe analizar detalladamente su comportamiento psciométrico[1] para así estimar la habilidad del evaluado. Comúnmente, este tipo de pruebas suele ser precedido por pruebas en papel y lápiz, y por tanto es indispensable garantizar comparabilidad entre los dos tipos de aplicación. Así mismo, se requiere considerar el balance de contenidos a través de los dominios a evaluar en cada prueba. En el Icfes, este se configura a partir de los Estándares Básicos de Competencias definidos por el Ministerio de Educación.

2. ¿Qué ventajas tiene esta forma de aplicación?

En CAT se genera una estimación provisional de la habilidad ($\hat{\theta}$) de cada evaluado con base en la respuesta dada, posterior a la administración de cada ítem. De manera que, para la población evaluada, se crean múltiples formas del examen; una de las diferencias entre la aplicación en papel y lápiz y CAT es que para el primero se preestablecen formas a partir de un banco de ítems, mientras que en CAT cada evaluado presenta un test individualizado (Wainer, Dorans, Eignor, Flaugher, Mislevy, Steinberg y Thissen, 2014).

Los beneficios de CAT han sido ampliamente documentados y tienen que ver con la precisión y eficiencia en la administración de ítems (Wainer et al., 2014; Magis, Yan, & Von Davier, 2017). Primero, se garantiza confiabilidad a nivel técnico y psicométrico a través de la idoneidad de la prueba, pues los ítems se enfocan en un rango de dificultad apropiado para cada evaluado. Segundo, este tipo de pruebas permite aumentar la seguridad, en la medida en que se controla la tasa de exposición de las preguntas. Tercero, y al igual que las demás pruebas en formato electrónico, no hay espacio para marcaciones incorrectas a causa de un borrado incompleto como sucede en papel y lápiz, y deja de lado la manipulación de hojas y cartillas. Cuarto, el aplicativo no permite múltiples marcaciones, ya que queda seleccionada la última opción señalada por el evaludo. Quinto, es posible generar un reporte inmediato de calificación, que como se ha visto en publicaciones previas, se realiza a través de TRI. Sexto, se administran menos preguntas que bajo la aplicación de papel y lápiz, puesto que los ítems se seleccionan de una manera más eficiente.

1. Bajo un modelo 3PL de Teoría Respuesta al Ítem, se realiza un análisis psicométrico de la discriminación, la dificultad y el azar de los ítems. Para profundizar sobre TRI y los parámetros del modelo refiérase a la serie 1 "¿Cómo se generan los puntajes en las pruebas Saber del ICFES?".

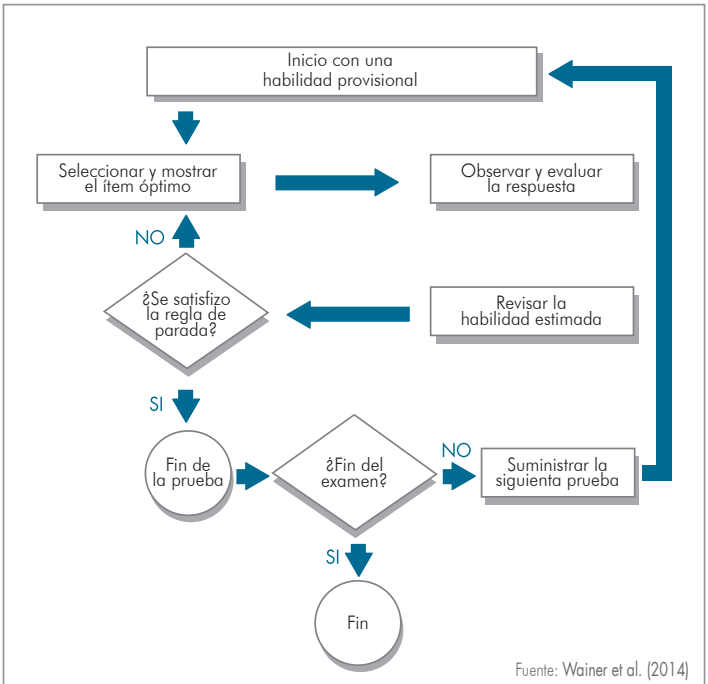
3. ¿En qué consiste?

La lógica que subyace detrás de la prueba adaptativa se presenta en la figura 1, donde se evidencia la existencia de criterios de (1) inicio, (2) selección de ítems y (3) parada. Se puede apreciar en el flujograma que la administración de ítems es un proceso iterativo que inicia con la selección y administración de ítems, seguido de la observación y la evaluación de la respuesta del evaluado al ítem, y una posterior estimación provisional de la habilidad ($\hat{\theta}$) que conduce a una decisión sobre la regla de parada. Este proceso se repite hasta que se satisfaga dicha regla de parada, y se estima la habilidad del evaluado.

Las pruebas adaptativas son una optimización de la versión computarizada equivalente a las pruebas de papel y lápiz, que consisten en seleccionar la siguiente pregunta de acuerdo a la habilidad actual del evaluado, iterativamente, durante la secuencia de preguntas del examen .

Dado que el aplicativo adaptativo solo se relaciona con la forma como se aplica un examen, es posible considerar la metodología de estimación de habilidades de los evaluados empleada en la aplicación de lápiz y papel: teoría respuesta al ítem (TRI). Bajo TRI se estima la relación entre la probabilidad de acertar un ítem y la habilidad de un evaluado, en términos de los parámetros de los ítems; y como en toda medición, surgen naturalmente errores de medición. Por ejemplo, si queremos medir la altura de una persona sin contar con una regla, nos enfrentamos a una situación en la cual no podemos tener con exactitud dicho valor, sino que debemos indagar y realizar preguntas relacionadas que nos aproximen a esa información. Por ejemplo, ¿al quedarte de pie en un microbús debes agacharte para ir cómodo? ¿Las bolsas de mercar rozan el piso cuando llevas las compras?, entre otras. Aquellas variables que no se pueden observar directamente se conocen como rasgos latentes, y a través de una indagación rigurosa es posible aproximarse a una medida cercana en cada medición. Volviendo a nuestro ejemplo, vamos a encontrar diferencias leves entre las medidas de altura por falta de una regla definida; estas variaciones se conocen como errores estándar, y es de interés que sean lo más pequeñas posibles con el fin de tener mayor precisión en la estimación. En modelos de TRI se emplea la función de información que depende de los parámetros de los ítems y permite calcular los errores estándar de medición (Demars, 2010).

FIGURA 1. Lógica de la prueba adaptativa



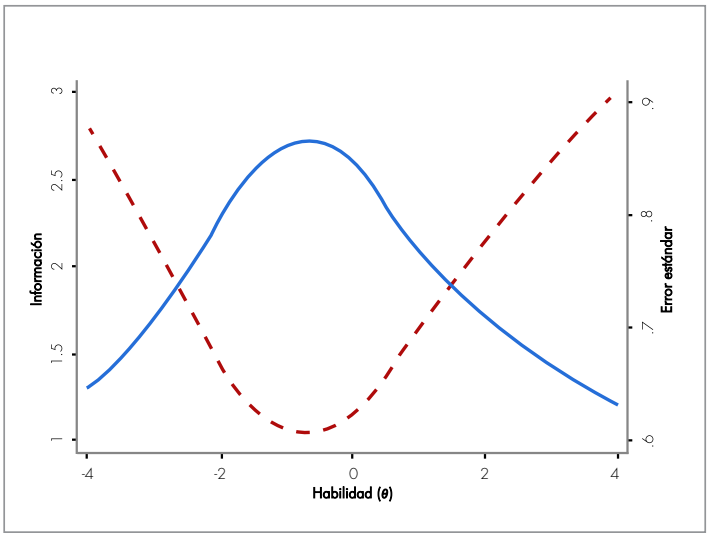
Fuente: Wainer et al. (2014)

¿Cómo es una función de información? En la figura 2 se presenta la interacción de dicha función con los errores estándar y la habilidad, y se observa que la función de información varía con la habilidad. Cabe notar que hay un punto de la función que brinda la máxima información para un rango particular de la habilidad. Como se observa, a mayor información es menor el error estándar de medición[2]. Este concepto de función de información es clave para el CAT, ya que es una herramienta útil para elegir los ítems siguientes para cada evaluado. Una de las ventajas de TRI es que cada ítem tiene su función de información, de tal

2. En términos técnicos el error estándar es el inverso de la raíz cuadrada de la información (Demars, 2010).

manera que la función de información del examen es la suma de las funciones de información de cada uno de los ítems en la prueba (Demars, 2010).

FIGURA 2. Función de información del examen



A continuación, se describen los tres momentos claves para el funcionamiento del CAT.

1. La elección del ítem inicial

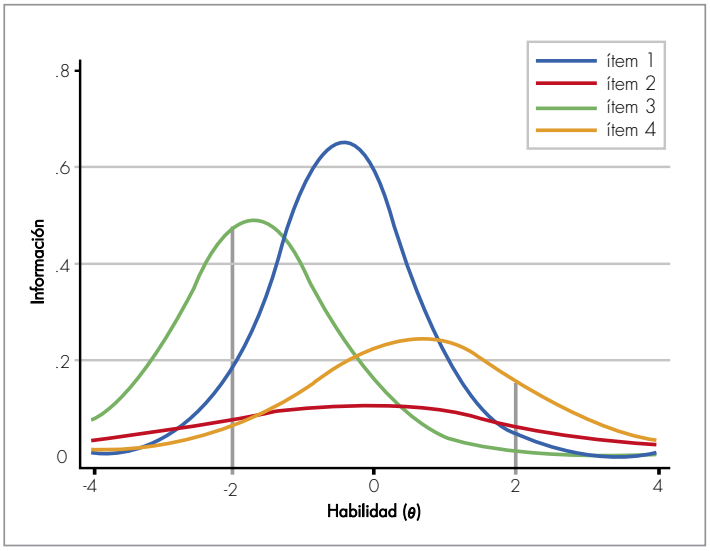
Hay diferentes criterios de inicio o de selección inicial. Por ejemplo, un criterio es asumir que la distribución de habilidades del evaluado es igual al promedio de la habilidad de los evaluados. En este caso, es razonable y óptimo pensar que la habilidad promedio estimada temporal es igual a la del promedio ($\hat{\theta} = \theta$), ya que después de unas respuestas, los evaluados se acercarán a ítems que son más informativos alrededor de su nivel de habilidad particular (Wainer et al., 2014). Se inicia con una distribución de habilidades normal, con media cero y desviación estándar 1. Así, se inicia con una pregunta de dificultad media, y dependiendo de su respuesta se aplica el siguiente ítem.

2. La elección del ítem siguiente después de ver la respuesta del anterior

Para medir efectivamente a todos los evaluados, el examen debe contar con un banco de ítems tal que permita al motor adaptativo aplicar ítems con dificultades cercanas al valor estimado provisional de habilidad, y recordemos que tanto la habilidad como la dificultad se miden bajo la misma métrica. En el proceso que se observa en la figura 1 tiene sentido, ya que se busca elegir en cada iteración el ítem que brinde el máximo de información, que resulta en la curva más informativa, y por tanto se ajuste cada vez mejor la habilidad provisional estimada $\hat{\theta}$ a la habilidad estimada θ. Así, la precisión del examen que mide a cualquier nivel de habilidad es proporcional al número de ítems cuyas dificultades coinciden con ese nivel.

En la figura 3, se presenta un ejemplo de la selección de un ítem para dos evaluados. En tal caso, vamos a suponer que el motor adaptativo va a elegir el segundo ítem del examen, y tenemos que la habilidad estimada provisional para la evaluada 1 es de $\hat{\theta} = 2$, y para la evaluada 2 es de $\hat{\theta} = -2$ según la respuesta al primer ítem. Bajo la lógica que describimos anteriormente, el motor adaptativo va a seleccionar aquel ítem que le brinde mayor información dada la estimación actual de la habilidad. Esto implica que para la evaluada con habilidad provisional $\hat{\theta} = 2$ el ítem que sigue va a ser el número 4, mientras que para la evaluada con habilidad provisional $\hat{\theta} = -2$, el ítem con la curva más informativa en ese rango de habilidad es el ítem 3. Tales ítems son los que brindan máxima información para cada una de las evaluadas, y como vemos son ítems distintos. Este proceso se realiza para seleccionar cada uno de los ítems siguientes, de acuerdo a la habilidad estimada de cada evaluado.

FIGURA 3. Función de información de los ítems



3. Criterio de parada

Hay varias formas de llegar a un criterio de parada. Uno de ellos tiene que ver con la longitud del examen, en el cual se establece el número de ítems que debe presentar cada evaluado. Al elegir el criterio de longitud de la prueba, es necesario simular el número de longitudes posibles, tales que permitan estimar con suficiente precisión la habilidad del evaluado y que cumplan a cabalidad las especificaciones de la prueba. Cabe señalar que la distribución de contenidos se debe contemplar en dicho análisis, para establecer el número óptimo de ítems. En ese caso, el número de ítems del examen administrado se selecciona proporcionalmente al número de ítems en los grupos de contenido definidos (Magis, Yan, & von Davier, 2017), por lo cual se requieren estrategias específicas. Así mismo, es posible tener un criterio de parada donde el motor adaptivo defina que el número de preguntas es suficiente para estimar con precisión la habilidad, o incluso bajo un tiempo definido para la aplicación del examen. Esos criterios dependen del evaluador y de las restricciones que presente.

Bibliografía

Demars, C. (2010). *Item Response Theory*. Oxford University Press.

Magis, D., Yan, D., & Von Davier, A. A. (2017). *Computerized Adaptative and Multistage Testing with R*. Springer International Publishing.

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 PRUEBAS ADAPTATIVAS POR COMPUTADOR (CAT) PARA LAS PRUEBAS SABER?

@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