La educación es de todos
Mineducación

SABER AL DETALLE

EDICIÓN
09
Bogotá D.C.
Marzo de 2022
ISSN: 2590 - 4663
Publicación Trimestral

¿CÓMO SE ANALIZAN LOS ÍTEMS DE LAS PRUEBAS SABER?

icfes mejor saber

Elaboración del documento
Nila Fernanda Amaya Melo
Astrid Julieth Betancourt Pineda
Patricia Escudero Montañez
Jenny Paola Martínez Fonseca
Carlos Arturo Parra Villamil
Mishell Marcela Ramos de la Hoz

DIseño y diagramación
Kevin Ostos Peñaloza

Bogotá D.C., marzo 2022

TÉRMINOS Y CONDICIONES DE USO PARA LAS PUBLICACIONES Y OBRAS QUE SON PROPIEDAD DEL ICFES

El Instituto Colombiano para la Evaluación de la Educación (Icfes) pone a disposición de la comunidad educativa, y del público en general, de forma gratuita y libre de cualquier cargo, un conjunto de publicaciones disponibles en su portal www.icfes.gov.co. Estos materiales y documentos están normados por la presente política, y se encuentran protegidos por derechos de propiedad intelectual y derechos de autor a favor del Icfes. Si tiene conocimiento de alguna utilización contraria a lo establecido en estas condiciones de uso, por favor infórmenos al correo prensaicfes@icfes.gov.co.

Queda prohibido el uso o publicación total o parcial de este material con fines de lucro. Únicamente está autorizado su uso para fines académicos e investigativos. Ninguna persona, natural o jurídica, nacional o internacional, podrá vender, distribuir, alquilar, reproducir, transformar*, promocionar o realizar acción alguna con la cual se lucre directa o indirectamente con este material. Esta publicación cuenta con el registro ISBN (International Standard Book Number o Número Normalizado Internacional para Libros), que facilita la identificación no solo de cada título, sino, también, de la autoría, la edición, el editor y el país en donde se edita.

En todo caso, cuando se haga uso parcial o total de los contenidos de esta publicación, el usuario deberá consignar o hacer referencia a los créditos institucionales del Icfes, respetando los derechos de cita. En otras palabras, se podrá hacer uso de esta publicación si dicho uso se contempla en los fines aquí previstos. Es posible, entonces, transcribir pasajes del texto si se cita siempre la fuente de autor. Por supuesto, estas citas no deberían ser excesivas ni frecuentes para que, así, no se considere una reproducción simulada y sustancial que redunde en perjuicio del Icfes.

Asimismo, los logotipos institucionales son marcas registradas y de propiedad exclusiva del Instituto Colombiano para la Evaluación de la Educación (Icfes). Por tanto, cuando su uso pueda causar confusión, los terceros no podrán usar las marcas de propiedad del Icfes con signos idénticos o similares respecto a cualquier producto o servicio prestado por esta entidad. En todo caso, queda prohibido su uso sin previa autorización expresa por parte del Icfes. La infracción de estos derechos se perseguirá civil y penalmente (en caso de que sea necesario), de acuerdo con las leyes nacionales y tratados internacionales aplicables.

El Icfes realizará cambios o revisiones periódicas a los presentes términos de uso y los actualizará en esta publicación.

* La transformación es la modificación de la obra a través de la creación de adaptaciones, traducciones, compilaciones, actualizaciones, revisiones, y, en general, cualquier modificación que se pueda realizar, haciendo que la nueva obra resultante se constituya en una obra derivada protegida por el derecho de autor, con la única diferencia, respecto de las obras originales, que aquellas requieren, para su realización, de la autorización expresa del autor o propietario para adaptar, traducir, compilar, etc. En este caso, el Icfes prohíbe la transformación de esta publicación. Términos y condiciones de uso para las publicaciones y obras que son propiedad del Icfes

¿CÓMO SE ANALIZAN LOS ÍTEMS DE LAS PRUEBAS SABER?

Las pruebas que desarrolla el Icfes son una herramienta de medición que permite estimar la habilidad de los evaluados en diferentes áreas del conocimiento de diferentes niveles educativos. Cada una de ellas se compone de agrupaciones de ítems, o preguntas, que se construyen siguiendo los lineamientos del Diseño Centrado en Evidencias[1] y, en conjunto, le apuntan a la medición de las competencias definidas de acuerdo con los Estándares Básicos de Competencias establecidos por el Ministerio de Educación Nacional. Un diseño y construcción adecuada de los ítems permite que el Icfes pueda cumplir su propósito de evaluar la educación que se ofrece en los distintos niveles educativos, tal como lo define la Ley 1324 del 2009. Como parte de las estrategias implementadas para asegurar la calidad de la evaluación realizada, tras cada aplicación de las pruebas, el Icfes realiza un conjunto de validaciones estadísticas sobre el comportamiento psicométrico de los ítems. En esta edición de Saber al Detalle se ahondará sobre esta estrategia, denominada análisis de ítems.

¿En qué consiste el análisis de ítems?

En la edición anterior[2] se presentaron los modelos de calificación de Teoría Clásica de los Test (TCT) y la Teoría de Respuesta al Ítem (TRI), y se mencionó que con ellos también se realiza la estimación de diversos estadísticos que se enmarcan en el proceso de *análisis de ítems* que se realiza previo a la calificación. Como su nombre lo indica, este proceso se centra en la revisión de tres aspectos: 1) un conjunto de indicadores que resumen el comportamiento psicométrico de cada ítem en función de la población evaluada, 2) el comportamiento de las curvas características de los ítems y 3) el posible comportamiento diferencial de los ítems entre aplicaciones.

Tras cada aplicación de los examenes Saber se realiza el análisis de ítems, en el cual, a partir de las respuestas de los evaluados a las preguntas en cada una de las pruebas, se estiman los parámetros de los ítems. Así, el proceso de análisis de ítems se lleva a cabo en tres etapas. En la primera de ellas, que se conoce como *análisis de ítem inicial*, se estiman los parámetros de los ítems utilizando únicamente las respuestas de la población en la última aplicación, luego se compara la estimación de estos parámetros con los obtenidos al utilizar los datos de aplicaciones anteriores; esto se realiza con el fin de detectar posibles cambios en el comportamiento

1. Para más información, refiérase a la Guía introductoria al diseño centrado en evidencias, disponible en: *https://www.icfes.gov.co/documents/20143/1500084/Guia+introductoria+al+Diseno+Centrado+en+Evidencias.pdf*

2. La edición anterior (Saber al Detalle, número 8), titulada "¿Cuáles son los modelos de calificación de las pruebas saber?", está disponible en: *https://www.icfes.gov.co/edicion-8-boletin-saber-al-detalle*

psicométrico del ítem a través del tiempo debido, entre otras razones, al cambio de las poblaciones a las que se aplicó, para lo cual se realiza un proceso llamado análisis de funcionamiento diferencial de los ítems. En la segunda etapa, que se conoce como *análisis de ítem de calibración*[3], se realiza un proceso de anclaje[4] de los parámetros de los ítems utilizando las estimaciones actualizadas a partir de la línea base de comparabilidad de resultados[5]. En la tercera, y última etapa, se realiza el *análisis de ítem de pilotos*. Aquí, se estiman los parámetros de los ítems nuevos, llamados pilotos, para estudiar su comportamiento y definir si sus propiedades psicométricas son adecuadas para incluirlos en el proceso de armado y calificación en exámenes posteriores.

A continuación, se describen los aspectos transversales del proceso de análisis de ítems.

¿Qué criterios poblacionales se tienen en cuenta para el análisis de ítems?

Para el análisis de ítems, se seleccionan únicamente los estudiantes inscritos por una institución educativa y que hayan estado presentes durante todas las sesiones del examen, que no sean sospechosos de copia, que hayan respondido más del 50 % de las preguntas de cada prueba, entre otros. Esta selección se realiza con el fin de tener la información lo más homogénea posible de la población objetivo del examen, es decir, de los estudiantes.

3. La calibración de ítems es el proceso por el cual se estiman los parámetros del modelo de calificación para cada uno de los ítems que conforman la prueba, con el fin de verificar que los ítems presenten un funcionamiento psicométrico óptimo.

4. Para más información sobre el proceso de anclaje, refiérase a la edición 3 del boletín Saber al Detalle, titulada "¿Qué garantiza la comparabilidad de los resultados en las pruebas Saber realizadas por el Icfes?" y disponible en: *https://www.icfes.gov.co/edicion-3-boletin-saber-al-detalle*

¿Por qué es necesario analizar los ítems?

El Icfes tiene un proceso de desarrollo cíclico de las pruebas que consiste en: el diseño, la elaboración, la aplicación, la lectura de respuestas, el análisis estadístico y la calificación. En particular, la etapa de análisis de ítems da información del comportamiento psicométrico de los ítems para que las personas que guían el desarrollo de las pruebas puedan tomar decisiones que contribuyan a mejorar la medición que realizan los ítems. Esta es una fuente importante de la validez de las pruebas y es imprescindible para garantizar la calidad de las evaluaciones.

La toma de decisiones depende de unos puntos de corte definidos previamente para cada una de las estadísticas obtenidas en el análisis de ítems. Estos puntos de corte se establecen con base en desarrollos teóricos y empíricos de evaluación que señalan qué valores reflejan un comportamiento atípico de los ítems para cada criterio. Bajo esta estructura, cuando un ítem presenta un comportamiento atípico, se le asigna una alerta y, por tanto, un ítem puede tener tantas alertas como indicadores se calculan. Dichas alertas permiten que los constructores y revisores analicen el contenido de los ítems, por ejemplo, su redacción, posibles problemas de impresión, la relevancia temática, cambios en el contexto cultural, entre otros (American Educational Research Association, 2014). Este análisis puede derivar en la exclusión de algunos ítems en el proceso de la calificación de la respectiva prueba o del conjunto de ítems que pueden emplearse en aplicaciones futuras. Lo anterior contribuye a fortalecer la evaluación, ya que al excluir los ítems que presentan un comportamiento atípico se identifican aspectos que afectan la estructura de las pruebas y se garantiza un proceso de evaluación de calidad.

5. De acuerdo con Resolución 268 de 2020 del Icfes, "por la cual se reglamentan las metodologías para la generación de resultados de los exámenes de Estado y se dictan otras disposiciones".

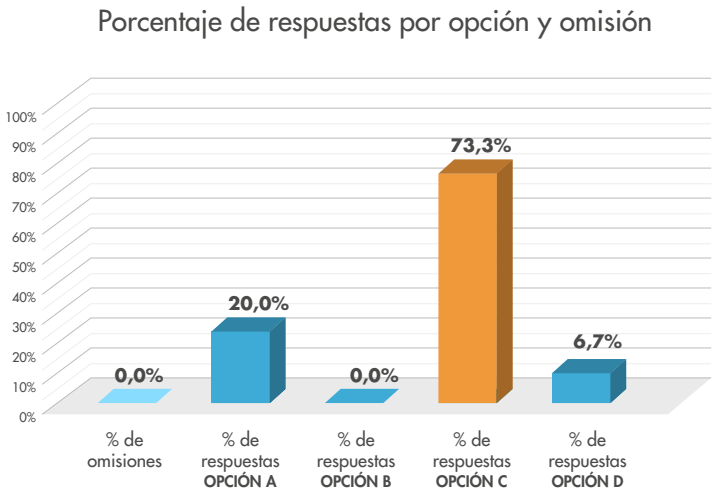
¿Qué estadísticas se analizan en un análisis de ítems general?

Un proceso de análisis de ítems se realiza en tres instancias. En la primera de ellas, se analizan los resultados descriptivos de las respuestas dadas por los evaluados al examen. Se calcula la proporción de evaluados que respondieron correctamente el ítem. Esta información viene acompañada de la distribución de los evaluados que elige cada opción de respuesta en el respectivo ítem, la cual permite analizar, por ejemplo, cuál fue la segunda opción de respuesta más seleccionada. Así mismo, se estima la proporción de evaluados que no dieron respuesta al ítem (omisiones) y la proporción de evaluados que seleccionaron más de una opción de respuesta en un mismo ítem (multimarca). A continuación, se presentan ejemplos de estas estadísticas para algunos ítems.

Como se observa en la **Figura 1a**, el 73,3% de la población seleccionó la opción C (barra de color amarillo), la cual corresponde a la respuesta correcta. Por otro lado, se observa que la opción D tiene un porcentaje de selección del 6,7% y no hubo evaluados que seleccionaron la opción B. Además, ninguno de los evaluados dejó el ítem sin responder (0 % de omisión). Esto podría sugerir que la opción C es fácilmente detectable como la opción correcta y que la opción B no provee información sobre la población, por tanto, se sugiere revisar todas las opciones del ítem.

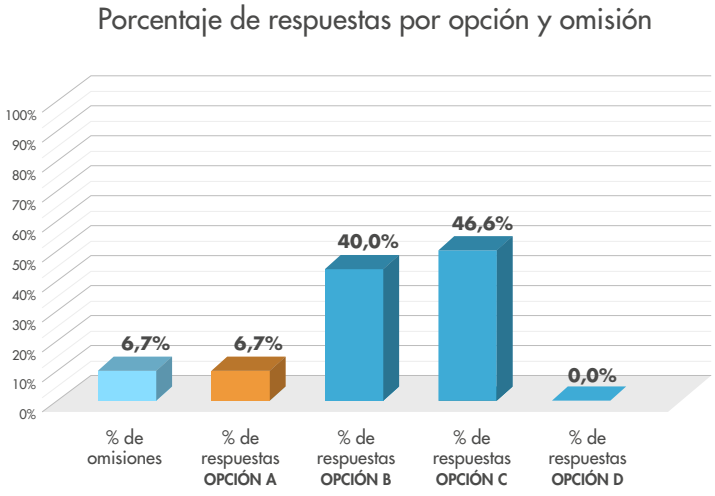
Por otro lado, la **Figura 1b** presenta un ejemplo en donde la opción C presenta un mayor porcentaje de respuestas (46,6%), seguido de la opción B con una proporción similar (40%). Mientras que la opción correcta, que en este caso es la A, fue seleccionada por un porcentaje bajo de evaluados (6,7%), respecto a las dos opciones mencionadas anteriormente. Por lo anterior es conveniente realizar una revisión del contenido del ítem con el fin de determinar si presenta errores en su contenido que genere que las opciones incorrectas sean llamativas.

FIGURA 1a. Distribución por opción de respuesta



Fuente: Icfes, Guía interpretación resultados 3° a 11° evaluar para avanzar, 2020

FIGURA 1b. Distribución por opción de respuesta



Fuente: Icfes, Guía interpretación resultados 3° a 11° evaluar para avanzar, 2020

En una segunda instancia, se hace un análisis bajo el modelo TCT en donde se calcula la correlación biserial, cuyos valores oscilan entre -1 y 1. Este estadístico expone el grado de asociación entre el acierto en el ítem y el total de respuestas correctas obtenidos por los evaluados en la población. En este caso, se calcula la correlación entre la selección de cada opción de respuesta y el puntaje de la prueba. Por lo tanto, es de esperar que una persona que obtiene un puntaje alto en la prueba tienda a responder correctamente los ítems, lo que daría como resultado correlaciones biseriales positivas y altas. También se espera que esta sea positiva y de mayor magnitud para la opción de respuesta correcta frente a las demás opciones de respuesta.

Adicionalmente, se calcula el coeficiente de confiabilidad Alpha de Cronbach, que indica qué tan consistente es la prueba, o de otra forma que tan relacionados están los ítems entre sí. Este estadístico oscila entre 0 y 1 y es uno de los indicadores de la precisión de los puntajes de la prueba. Dentro del Instituto se cuenta con los siguientes valores de referencia para este coeficiente:

- Confiabilidad menor a 0,6 es baja y podría sugerir dificultades para la estimación de la habilidad de los evaluados.

- Confiabilidad entre 0,6 y 0,75 es baja, pero aceptable, lo que sugiere que, aunque hay aspectos para mejorar, permite estimar adecuadamente la habilidad de los evaluados.

- Confiabilidad entre 0,75 y 0,95 es óptima y podría sugerir que las medidas de las pruebas cuentan con un buen nivel de precisión en la estimación de la habilidad.

- Confiabilidad mayor a 0,95 sugiere redundancia en los ítems utilizados para estimar la habilidad de los evaluados.

En una tercera instancia se hace un análisis que se basa en los resultados de la TRI. Es importante recordar que a partir de esta teoría es posible estimar los parámetros de los ítems de discriminación (a), dificultad (b) y pseudo-azar (c), dependiendo del modelo acogido[6]. El primer parámetro está asociado con la capacidad del ítem de diferenciar entre los evaluados con habilidades altas de los evaluados con habilidades bajas; el segundo, con el nivel necesario de habilidad para contestar correctamente el ítem alrededor del valor en que mejor discrimina; y el tercero, con la probabilidad de contestar correctamente el ítem sin tener el nivel de habilidad necesario para acertarlo. Una vez estimados estos parámetros, se calculan los valores de las habilidades[7] para cada ítem y la probabilidad de escoger una opción de respuesta dependiendo del nivel de habilidad. En la **Figura 2**, que ilustra este análisis, se observa la habilidad y la proporción de evaluados que seleccionó cada opción de respuesta de un ítem. En este caso, la curva para la opción correcta B ilustra que a mayor habilidad hay una proporción más alta de evaluados que elige dicha opción. Por otra parte, las curvas de las opciones de respuesta incorrectas A, C y D sugieren que los evaluados con menor habilidad suelen

6. Para más información sobre los parámetros contemplados por los modelos de calificación de TRI, refiérase a la edición 8 del boletín Saber al Detalle, titulada "¿Cuáles son los modelos de calificación de las pruebas Saber" y disponible en *https://www.icfes.gov.co/edicion-8-boletin-saber-al-detalle*

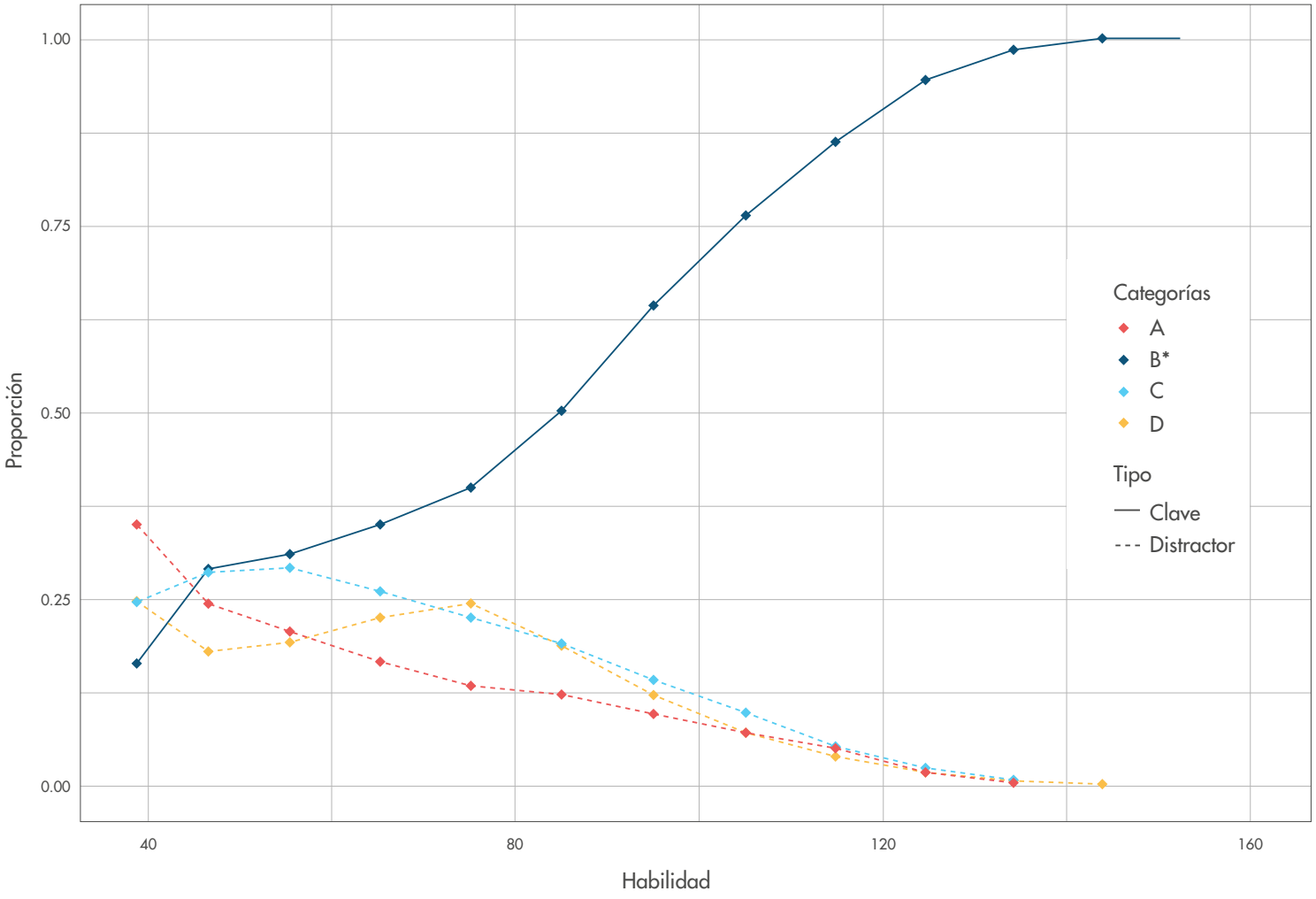
7. Para más información sobre la estimación de la dificultad, la discriminación y la probabilidad de acierto casual de cada ítem a través de la función de máxima de verosimilitud de los datos, refiérase a la edición 1 del boletín Saber al Detalle, titulada "¿Cómo se generan los puntajes en las pruebas Saber del Icfes?" y disponible en *https://www.icfes.gov.co/edicion-1-boletin-saber-al-detalle*

seleccionar más estas opciones, mientras que las personas de mayor habilidad no lo hacen.

Siguiendo el ejemplo anterior, se observa que los evaluados con habilidades bajas tienden a responder cualquier opción y, a medida que aumenta la habilidad del evaluado, aumenta la probabilidad de seleccionar la opción correcta (en este caso, la opción B).

Por ejemplo, a partir de una habilidad cercana a 60, es más probable que los evaluados elijan la opción de respuesta correcta.

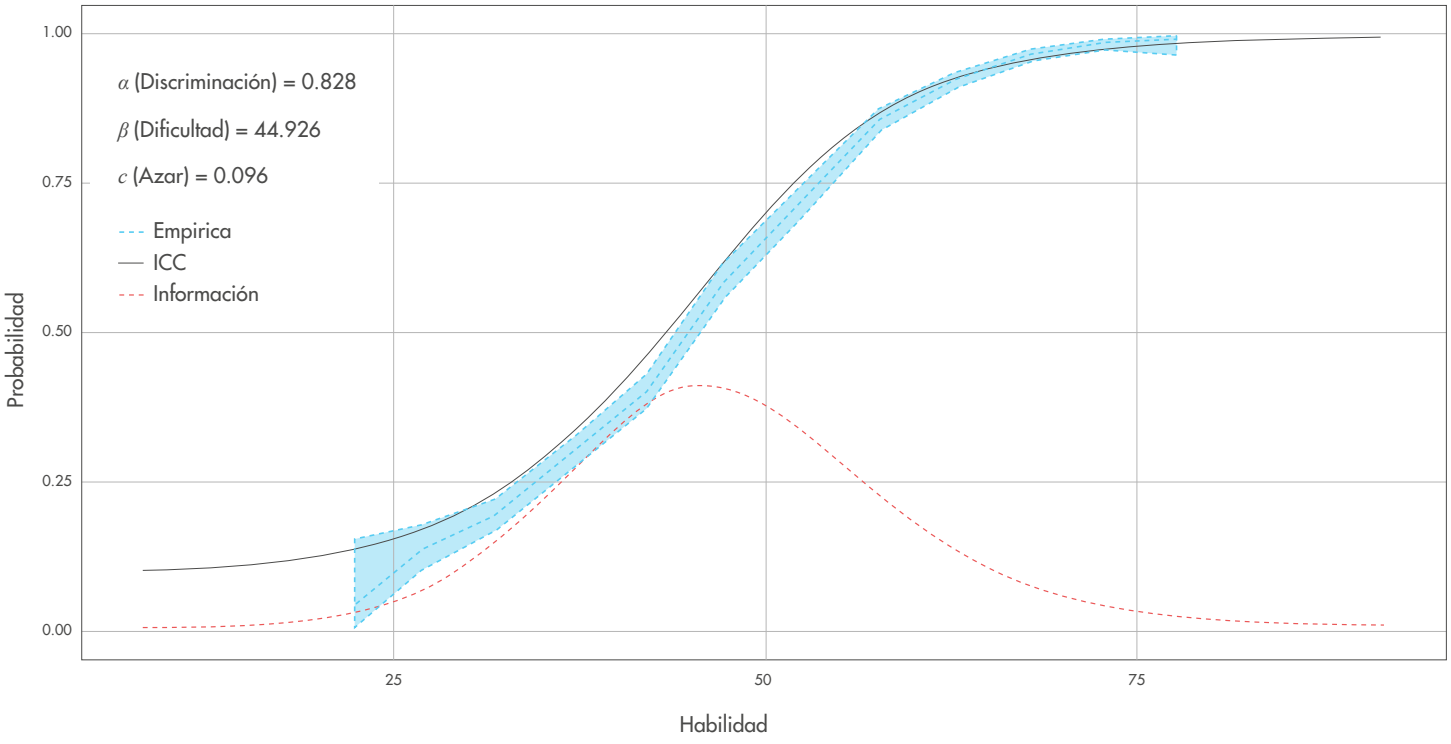
FIGURA 2. Curva del ítem por opción de respuesta



Fuente: Icfes, 2020

Con base en la estimación de los parámetros de los ítems libres[8] y de la habilidad se construyen dos curvas características del ítem (CCI)[9]. En la **Figura 3** se observa tanto la curva empírica (en color azul) como la curva ajustada por el modelo TRI de 3 parámetros (en color negro), así como la curva de información (en color rojo) que tiene una relación con los errores estándar de medición de las habilidades[10]. La primera curva nos da información sobre las puntuaciones observadas y la segunda sobre las puntuaciones esperadas teóricamente dadas las estimaciones de los parámetros del ítem. Como se observa, las dos curvas tienen un comportamiento similar y, por tanto, no hay motivos para sospechar que haya afectación en la estimación de los parámetros de los ítems por evaluados con comportamientos atípicos. Así mismo, se observa que el punto donde se obtienen mediciones más precisas del ítem es para habilidades cercanas a 45, que corresponden al punto más alto de la distribución de la curva de información. Por otra parte, habrá mayor probabilidad de error en la medición del ítem para los niveles de habilidad lejanos a este valor.

FIGURA 3. Curva característica del ítem y curva de información



Fuente: Icfes, 2020

8. Los ítems libres se definen como aquellos que sus parámetros no se han estimado con anterioridad. Por otra parte, los ítems que cuentan con esta información se conocen como ítems históricos.

9. Para más información sobre la curva de información del ítem, refiérase a la edición 5 de Saber al Detalle, titulada "¿En qué consiste la aplicación de pruebas adaptativas por computador (CAT) para las pruebas Saber?" y disponible en *https://www.icfes.gov.co/edicion-5-boletin-saber-al-detalle*

10. Cuanto más alta la curva de información, menor error de medición en la habilidad. Así, el segmento de habilidad con mayor información es el segmento que con menor error de medición.

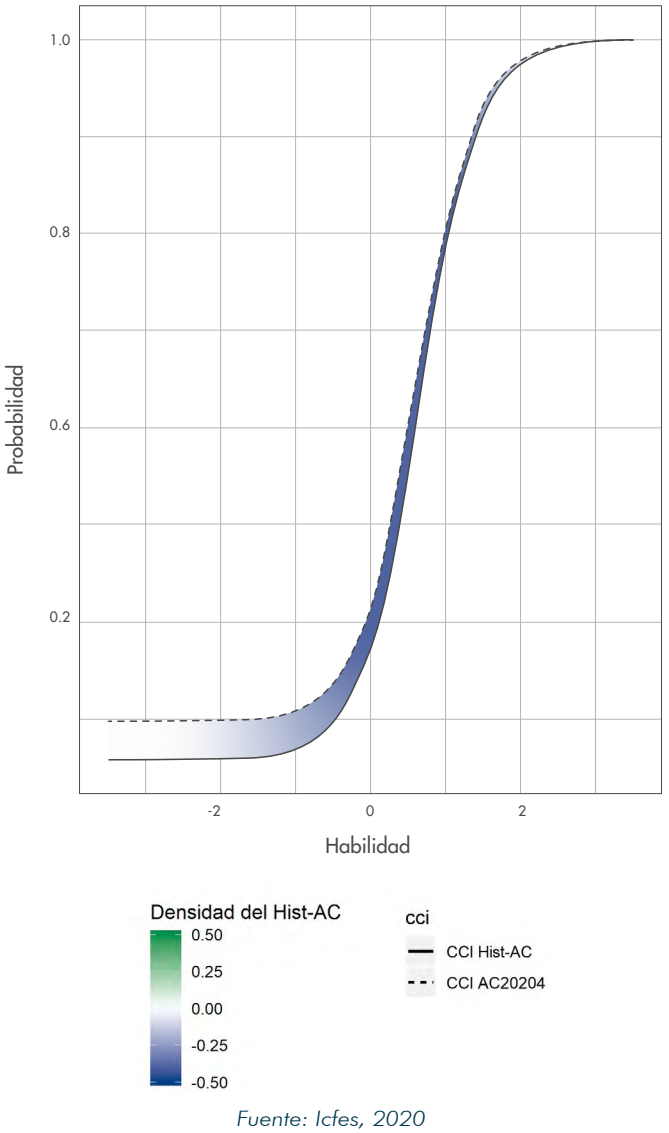
¿Cómo se comporta el ítem entre distintas aplicaciones?

Una vez se realiza este análisis, se busca que la medición de la habilidad de los evaluados sea comparable con las aplicaciones previas, para lo cual se realiza una comparación del comportamiento de los ítems en la aplicación actual frente a su comportamiento histórico. En esta comparación se analiza qué tan diferente es la probabilidad de acertar el ítem con un mismo nivel de habilidad frente al mismo en una aplicación pasada, por medio de las diferencias en los parámetros del ítem del periodo analizado frente a periodos anteriores. Si llega a existir alguna diferencia en esa comparación se afirma que se presenta un Funcionamiento Diferencial del Ítem (DIF, por sus siglas en inglés).

En términos prácticos, el análisis de DIF que se desarrolla en el Icfes busca generar una alerta respecto a diferencias grandes en las curvas características de los ítems, pues dichas curvas son la representación gráfica de la función que relaciona la probabilidad de acertar el ítem en términos de la habilidad y de los parámetros del ítem que definen la escala de calificación de la prueba. La estadística que se emplea para ello es el Índice No Compensatorio de DIF (NCDIF), el cual mide las distancias de las curvas características de los ítems, ponderando estas diferencias por la distribución de la habilidad, en este sentido, este análisis está relacionado con los parámetros específicos de los ítems como dificultad, discriminación y azar (Bolt, 2002; Oshima, Raju, & Nanda, 2006).

La **figura 4** presenta un ejemplo de una gráfica asociada al análisis de DIF. En ella se observan dos curvas características del ítem: una relacionada con los parámetros históricos (curva continua) y otra con los parámetros del periodo analizado (curva punteada). Se observa que hay una diferencia en la probabilidad de acierto a lo largo de la distribución de habilidad a favor de la CCI del periodo analizado, y esta diferencia se reduce en los niveles de habilidad más altos.

FIGURA 4. Curvas características del ítem[11]



Fuente: Icfes, 2020

11. En esta figura, la habilidad está expresada en logits.

Posteriormente, se aplica un método estadístico de equiparación[12] que permite definir una métrica común entre aplicaciones que pueden tener diferencias en los parámetros de los ítems, en conjunto, los pasos anteriores permiten realizar comparaciones directas. Teniendo en cuenta que el grupo de evaluados varía entre aplicaciones, la estimación de los parámetros de los ítems puede presentar diferencias. Una vez realizado dicho procedimiento, se repite el análisis de ítem ilustrado anteriormente, para la etapa de calibración y pilotos.

Finalmente, se realiza un comité en el cual se revisa la pertinencia estadística y conceptual de cada ítem que presenta un comportamiento atípico y como resultado de la revisión realizada en esta instancia, un ítem puede: *conservarse* para su uso en la calificación de la prueba o *eliminarse* de esta evaluación.

12. Para más información sobre los métodos de equiparación, refiérase a la edición 3 de Saber al Detalle, titulada "¿Qué garantiza la comparabilidad de los resultados en las pruebas Saber realizadas por el Icfes?" y disponible en *https://www.icfes.gov.co/edicion-3-boletin-saber-al-detalle*

Bibliografía

American Educational Research Association, American Psychological Association & National Council for Measurement in Education [AERA, APA & NCME] (2014). *The Standards for Educational and Psychological Testing*. Washington, D.C.: AERA.

Bolt, D. (2002). *A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods.* Applied Measurement in Education, 15, 113-141.

Congreso de la República. (13 de julio de 2009). *Ley 1324 de 2009*. DO: 47.409.

Guilford, J. P. (1975). *Factors and factors of personality. Psychological Bulletin*, 82(5), 802–814.

Icfes (2020). *Esquema de análisis - Módulo análisis de ítem, análisis univariados, TCT y TRI.* Subdirección de Estadísticas.

Icfes (2020). *Guía de interpretación de resultados 3° a 11° evaluar para avanzar.* Instituto Colombiano para la Evaluación de la Educación.

Mislevy, R. et al. (2003). *A brief introduction to evidence-centered design.* Educational Testing Service, Princeton, NJ.

Oshima, T., Raju, N. y Nanda, A. (2006). *A new method for assessing the statistical sign (American Educational Research Association, 2014)ificance in the differential functioning of items and tests (DFIT) framework.* Journal of Educational Measurement, 1-17.

SABER AL DETALLE

¿CÓMO SE ANALIZAN LOS ÍTEMS DE 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