MINISTERIO DE EDUCACIÓN NACIONAL

Informe técnico sobre el examen de Estado

en tiempos de la COVID-19

Saber 11

Subdirección de Estadísticas
Dirección de Evaluación

Bogotá, diciembre de 2022

Examen Saber 11° | icfes

MINISTERIO DE EDUCACIÓN NACIONAL

Elaboración del documento
John Alexander Calderón Rodríguez
Luis Adrián Quintero Sarmiento
Karen Rosana Córdoba Perozo
Nelson Andrés Rodríguez Rivera
Carlos Arturo Parra Villamil
Nila Fernanda Amaya Melo

DIseño y diagramación
Kevin Ostos Peñaloza

Fotografía portada
Flickr Ministerio de Educación
https://www.flickr.com/photos/mineducacion/46038072082/

ISBN: 978-958-11-1009-4

Bogotá D.C., diciembre 2022

Informe técnico sobre el examen de Estado

en tiempos de la COVID-19

Saber 11

Examen Saber 11° | icfes

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[1], 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.

1 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

Tabla de **contenido**

Índice de **figuras**

Índice de **tablas**

Introducción

El examen Saber 11 se aplica de manera periódica, principalmente a los estudiantes que están a punto de finalizar la educación media en Colombia. Sin embargo, con la emergencia sanitaria asociada a la COVID-19, el Instituto requirió de la implementación de un conjunto de ajustes sobre el diseño de la prueba, esto con de fin de minimizar el riesgo de contagio de los estudiantes. No obstante, antes de implementar las modificaciones, era necesario estimar el efecto que podrían acarrear sobre las propiedades psicométricas de la prueba.

En este sentido, desde la Dirección de Evaluación del Instituto Colombiano para la Evaluación de la Educación (Icfes) se plantearon una serie de escenarios para tomar decisiones sobre la mejor manera de mantener la aplicación del examen durante la situación de pandemia.

Atado a esto, desde la Subdirección de Estadísticas se propusieron una serie de indicadores sobre los cuales se podría simular el efecto de las decisiones a tomar. En esta línea, el presente documento parte desde la importancia de la aplicación del examen, posteriormente describe los ajustes implementados, los escenarios simulados, los resultados de las simulaciones y finalmente, presenta las principales conclusiones sobre el efecto de las decisiones tomadas.

01.

Sobre el examen **Saber 11**

El Icfes tiene como misión evaluar el cumplimiento de los objetivos planteados para el sector educativo a través de los exámenes de estado para los distintos niveles de educación (Ley 1324, 2009). Para el caso de la educación media, el Instituto ha desarrollado el examen Saber 11, el cual es una evaluación estandarizada que brinda información frente a las competencias que han desarrollado los evaluados a lo largo de su proceso educativo. Este examen es presentado por estudiantes que se encuentran finalizando el grado undécimo, así como también, quienes han obtenido el título de bachiller o superado el examen de validación del bachillerato.

Teniendo en cuenta lo establecido en el Decreto 869 de 2010, el examen Saber 11 tiene como objetivo comprobar el grado de desarrollo de las competencias de los estudiantes que están por finalizar el grado undécimo de la educación media, así como monitorear la calidad de la educación de los establecimientos educativos del país, con fundamento en los estándares básicos de competencias y los referentes de calidad emitidos por el Ministerio de Educación Nacional (MEN). Además, el examen sirve como referente estratégico para el establecimiento de políticas educativas nacionales, territoriales o institucionales.

El examen Saber 11 está compuesto por las pruebas de Lectura crítica, Matemáticas, Sociales y ciudadanas, Ciencias naturales e Inglés, las cuales permiten evaluar el desempeño de los estudiantes en las competencias definidas por el MEN en los estándares básicos. Adicionalmente, durante el examen se aplica un cuestionario que recoge información sobre las características sociodemográficas de las personas evaluadas. La **Figura 1** presenta la estructura para la aplicación del examen hasta el año 2019, en donde este se realizaba en dos sesiones con una duración de 4 horas y 30 minutos cada una.

Figura 1. Organización de la aplicación del examen Saber 11

		Número de preguntas	Total de preguntas por sesión	Tiempo por sesión
Primera sesión	Matemáticas ●	25	131	4 h y 30 min
	Lectura Crítica	41		
	Sociales y Ciudadanas ●	25		
	Ciencias Naturales ●	29		
	Cuestionario socioeconómico ●	11		
Segunda sesión	Sociales y Ciudadanas ●	25	147	4 h y 30 min
	Matemáticas ●	25		
	Ciencias Naturales ●	29		
	Inglés	55		
	Cuestionario socioeconómico ●	13		

● Parte 1 ● Parte 2

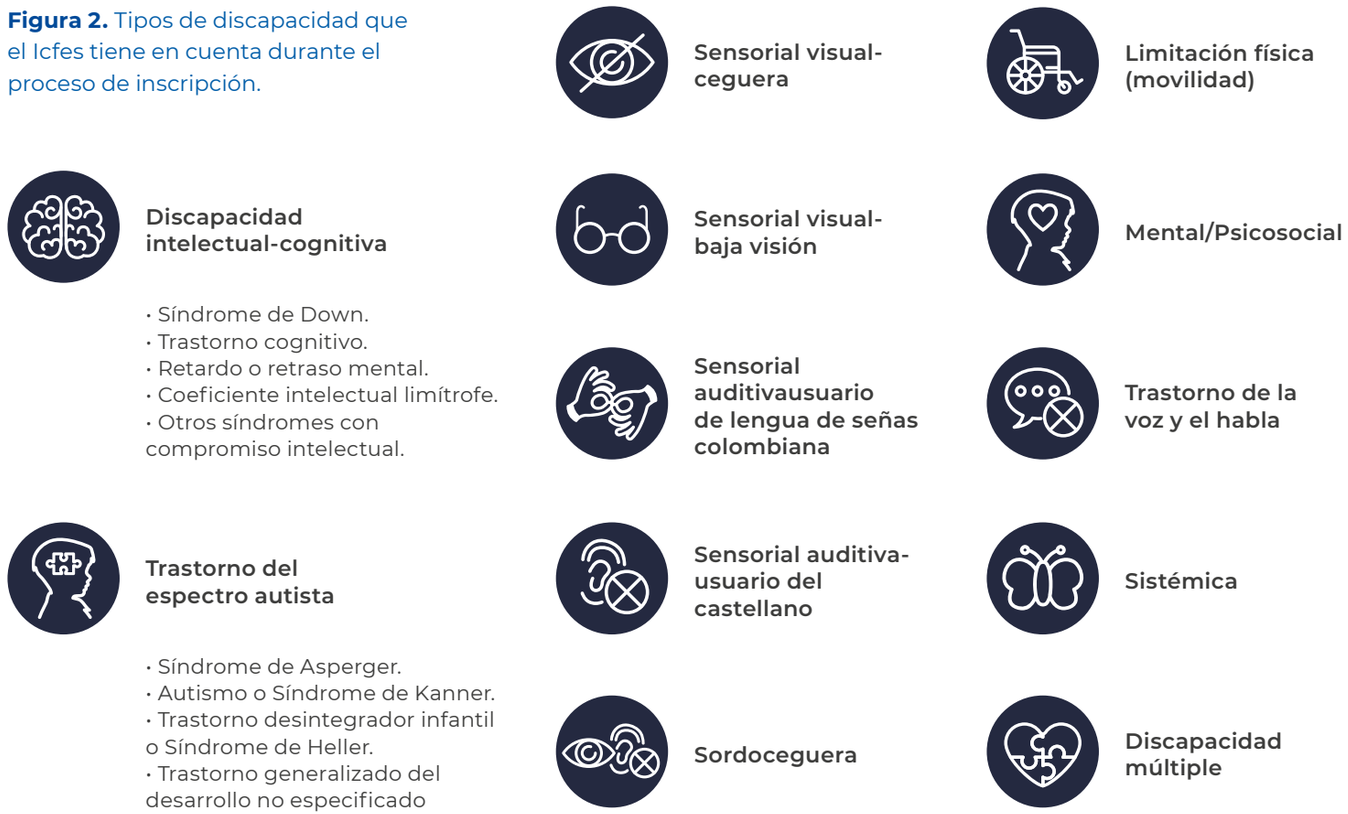
Nota: Figura tomada de la guía de orientación 2021-1.

La anterior organización del examen aplica para los estudiantes con discapacidad motora y sin discapacidad. Para las personas con algún tipo de discapacidad diferente a la motriz que se presentan al examen, el Icfes ha creado versiones de los cuadernillos con ajustes razonables específicamente para esta población con el fin de "(...) garantizar una adecuada y equitativa evaluación del desarrollo de competencias de las personas con discapacidad" (Resolución 675, 2019). Estos cuadernillos cuentan con una menor cantidad de preguntas y, además, el Icfes brinda apoyos a estas personas para que realicen su evaluación. Para esto, el instituto recoge información durante el proceso de inscripción frente al tipo de discapacidad que reportan los evaluados (Ver **Figura 2**).

Los evaluados que reportan algún tipo de discapacidad pueden elegir entre el cuadernillo con ajustes razonables o el estándar que es aplicado a la población general. Adicionalmente, esta población puede elegir si presenta la prueba de Inglés o no. Estas opciones dejan a los evaluados con discapacidad la posibilidad de presentar la prueba con ajustes, o, si ellos lo desean, la prueba estándar presentada por la población sin discapacidad.

Durante la pandemia fue necesario realizar algunos ajustes para el examen Saber 11, de manera que se pudiera aplicar teniendo en cuenta las restricciones generadas por la COVID-19. Dado lo anterior, el presente documento presenta los cambios que se realizaron y la metodología utilizada para garantizar que los puntajes de las pruebas se mantuvieran comparables con la versión de Saber 11 aplicada antes de la pandemia.

Figura 2. Tipos de discapacidad que el Icfes tiene en cuenta durante el proceso de inscripción.



Nota: Imagen adaptada a partir de la guía de orientación Saber 11 2021-2.

02.

Ajustes en el examen Saber 11 **asociados a la emergencia sanitaria - COVID-19**

Examen Saber 11°

En el primer semestre del 2020 se tenía planeada la aplicación de Saber 11 para calendario B. Por lo tanto, la logística de aplicación ya estaba preparada y los cuadernillos estaban impresos en el momento en que se declaró el estado de emergencia por la COVID-19 en marzo de 2020. Con el fin de garantizar la aplicación de la prueba a las personas que se habían inscrito, esta se aplazó y se modificaron los aspectos logísticos en cuanto al número de colegios donde se llevaría a cabo la aplicación. Más específicamente, se incrementó el número de sedes donde se aplicaría la prueba, de manera que se tuvieran menos estudiantes por salón y se pudiera garantizar el distanciamiento social durante el desarrollo de la prueba. Esto fue posible gracias a que el número de estudiantes en calendario B es relativamente pequeño y se logró conseguir la cantidad de colegios requeridos para esto.

En cuanto a la aplicación para calendario A durante el año 2020, fue necesario tomar algunas medidas debido al gran número de estudiantes que presentarían la prueba. Para tener mayor distanciamiento durante la aplicación, fue necesario reducir el tiempo de la prueba, pasando de dos sesiones a una, de manera que la mitad de los estudiantes fueran evaluados en la mañana y la otra mitad en la tarde. Para lograr disminuir el tiempo de aplicación de la prueba, fue necesario reducir la cantidad de preguntas que componen el examen y no se incluyeron los ítems piloto. Con estas modificaciones, el tiempo de aplicación de la prueba para cada estudiante fue de 5 horas y 30 minutos (Icfes, 2020).

Para brindar un panorama frente a las últimas aplicaciones de las pruebas Saber 11, en la **Tabla 1** se presenta el número de personas evaluadas según calendario y tipo de inscripción (Estudiante o Individual[1]), entre el 2018 y 2020. Con estos datos se evidencia que el número de evaluados disminuyó durante el año 2020 de pandemia. Además, en las segundas aplicaciones del año, dirigidas principalmente a los estudiantes de calendario A, el número de personas inscritas como estudiantes es mayor que el número de inscritos como individuales, a diferencia del calendario B, en el cual se invierte la relación.

Tabla 1. Número de evaluados según: año, tipo de inscripción y calendario del examen Saber 11 en el periodo 2018 - 2020

Año	Segunda aplicación (calendario A)		Primera aplicación (calendario B)	
	Estudiante	Individual	Estudiante	Individual
2018	555.317	53.817	19.798	46.056
2019	554.662	60.018	21.342	44.762
2020	514.152	42.747	15.435	30.786

1 Las personas inscritas como individuales son quienes presentan el examen después de obtener su título como bachiller. (Icfes, 2020)

Al relacionar estas tendencias con las medidas tomadas en el marco de la emergencia sanitaria, es relevante mencionar que, desde el punto de vista logístico, el Instituto ideó una serie de estrategias para evitar las aglomeraciones en las instituciones educativas, con el apoyo del personal de los sitios en los cuales se aplicó el examen. Además de reducir el número de estudiantes por salón, se dispuso gel antibacterial y lavaderos de manos. Así mismo, se incrementó el número de veces que se realizaba aseo en las sedes para seguir las recomendaciones de bioseguridad durante la pandemia.

Por otro lado, sumados a las decisiones logísticas, desde la Dirección de Evaluación se incluyeron algunos ajustes con respecto a la estructura de la prueba como se describe a continuación.

2.1. Ajustes referidos al número de ítems

Un primer ajuste en la estructura de la prueba para la aplicación 2020-2 tiene que ver con los ítems piloto. En la Tablas 2 y 3 se presentan los conteos del número de ítems pilotos presentados por cada evaluado para las pruebas de Saber 11 en 2019-2 y 2020-1. Para un periodo específico, un evaluado presenta en total 33 ítems pilotos y se estima que el tiempo invertido por un evaluado para contestar cada ítem es de aproximadamente 2 minutos en promedio. Dado lo anterior, la reducción en tiempos al eliminar los ítems piloto para la aplicación de 2020-2 durante la pandemia corresponde a 66 minutos aproximadamente, lo cual es un tiempo bastante considerable.

Examen Saber 11°

03.

Metodología para la implementación de ajustes en el examen Saber 11 **debido a la emergencia sanitaria**

Además de eliminar los ítems piloto de la prueba en 2020-2, fue necesario disminuir el número de ítems no piloto con el fin de recortar la prueba a una sola sesión. A continuación, se discuten los aspectos metodológicos tenidos en cuenta para tomar la decisión de disminuir la cantidad de ítems en la prueba durante la aplicación de Saber 11 de 2020-2 debido a la emergencia sanitaria. Para esto, se presentan algunas consideraciones sobre el diseño de armado de las pruebas, la metodología de calificación y el error de medición en la estimación de los puntajes de los estudiantes. Asimismo, se discuten los aspectos técnicos tenidos en cuenta para asegurar la comparabilidad de la prueba al haber realizado estas modificaciones de aplicación.

3.1. Armado Saber 11° - BIBs

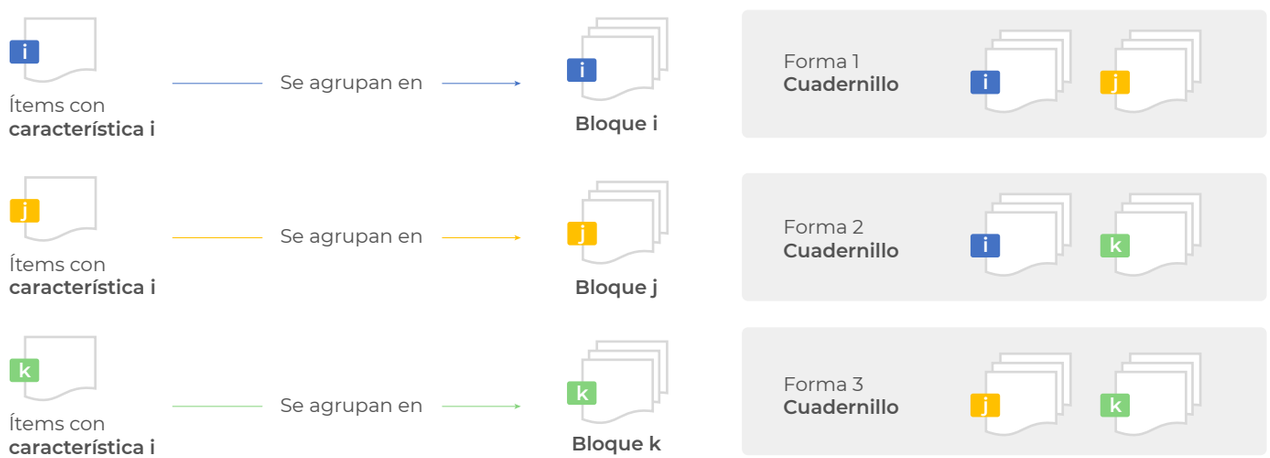
Al interior del Instituto se utiliza una metodología específica de armado denominada Diseño por bloques incompletos balanceados (BIBs), en la cual, los ítems se organizan en subconjuntos de estos que se denominan bloques, los cuales están compuestos por el mismo número de ítems, sin embargo, estos últimos son distintos en cada bloque. Otra característica de este diseño es que los bloques se combinan entre sí para construir formas de prueba. En esta misma línea, también se dice que en este diseño los bloques están incompletos porque una forma no contiene todos los bloques de la prueba. Finalmente, sobre este diseño también se afirma que está balanceado porque cada bloque y las combinaciones por pares de bloques se repite el mismo número de veces a lo largo de todas las formas dispuestas para la evaluación, respectivamente.

En el caso del Icfes, la selección de los ítems que componen cada bloque está alineada con la ponderación de las especificaciones de prueba y algunos criterios estadísticos preestablecidos que están asociados a las características de los ítems (por ejemplo: promedio de dificultad, desviación estándar de la dificultad, entre otros), para garantizar la equivalencia entre formas.

Esta metodología se emplea con el fin de garantizar que el instrumento de medición cumpla con ciertas cualidades psicométricas que permita evaluar los contenidos y dominios definidos en el marco conceptual de la prueba. A continuación, se ilustra el funcionamiento de esta metodología con un ejemplo práctico en el cual se pretende mostrar la complejidad y las implicaciones del proceso de armado sobre la aplicación del examen.

Para empezar, es necesario mencionar que cada cuadernillo presentado por una persona se conoce al interior del Instituto como forma. En una misma aplicación, para cada una de las cinco pruebas se generan varias formas, las cuales incluyen ítems con dificultades similares que dan cuenta del mismo rango de habilidad de los evaluados. A su vez, una forma está compuesta por varios subconjuntos de ítems llamados bloques. En general, la manera en la que se arman los cuadernillos o formas garantiza que cada bloque se repita un mismo número de veces en las diferentes versiones que se le pueden asignar a un estudiante (Ver **Figura 3**).

Figura 3. Esquema de ejemplo de agrupación de ítems en una prueba que se arma con dos bloques.



Es pertinente mencionar que, en las pruebas de Competencias ciudadanas, Lectura crítica, Ciencias naturales y Matemáticas, se utilizan generalmente cuatro bloques (conjuntos de ítems) para conformar el cuadernillo que se le entrega a cada estudiante. Con el fin de reducir el número de preguntas y así exponer menos tiempo a las personas evaluadas al posible contagio de la COVID-19, se optó por eliminar uno de los bloques para la aplicación de 2020-2.

Para tomar la decisión de cuál bloque eliminar, garantizando que la calidad de la medición se viera lo menos afectada posible tras el acortamiento de la prueba, se realizaron algunos ejercicios de simulación con los datos de 2018 eliminando uno de los cuatro bloques que presentaba cada estudiante por prueba. Con base en lo anterior, se seleccionó el bloque que presentaba los valores más bajos en el estadístico de información del test (De Ayala, 2018) para estimar la habilidad de los estudiantes. A continuación, se presentan los elementos necesarios para el cálculo de la función de información de los ítems y de los bloques en la prueba.

3.2. Modelos de calificación 3PL

Las pruebas Saber se califican con base en modelos de Teoría de Respuesta al Ítem (TRI), en los cuales se asume que la probabilidad de responder correctamente a un ítem es una función logística que depende de la habilidad del evaluado y de ciertos parámetros de cada ítem. En particular, la prueba Saber 11 se califica a partir del modelo TRI de tres parámetros (3PL), de manera que para cada ítem i se tiene un parámetro de dificultad (b_i), un parámetro de discriminación (a_i), y un parámetro de pseudo-azar (c_i). Bajo este modelo, la probabilidad de que el estudiante j conteste correctamente al ítem i se define como

$$P(U_{ij}=1 \mid \theta_j, a_i, b_i, c_i) = c_i + (1-c_i)\frac{e^{a_i(\theta_j - b_i)}}{1+e^{a_i(\theta_j - b_i)}}$$

Donde θ_j es la habilidad del estudiante j, a_i se interpreta como la pendiente máxima de la curva característica de cada ítem y el parámetro c_i es la probabilidad de acertar el ítem para un estudiante con habilidad cercana a -∞. Al utilizar modelos TRI, el interés se centra en estimar la habilidad de los evaluados principalmente, pero también es importante estimar los valores de los parámetros de los ítems (a_i, b_i y c_i) para determinar sus características psicométricas. Con base en estos parámetros estimados, se puede aproximar la precisión en la estimación de los puntajes de los evaluados, como se muestra a continuación.

3.3. Precisión de estimación y curvas de información

Además de estimar la habilidad θ de los evaluados, es importante evaluar la precisión de dichas estimaciones a través del error estándar de medición (EE), basado en el modelo TRI. Errores pequeños indican bajos niveles de variabilidad del estimador, y, por lo tanto, niveles altos de precisión en la estimación. Siguiendo a De Ayala (2018), el cálculo del EE de θ se hace a partir del inverso de la raíz cuadrada de la función de información del test (FIT), es decir,

$$EE(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

Donde $I(\theta)$ es el valor de la función de información para una habilidad θ dada y se calcula a partir del modelo de calificación. Al tener mayor información FIT, se tiene un error menor y una precisión más alta para la estimación de la habilidad. La FIT se calcula como la suma de la función de información de los ítems que componen el test:

$$I(\theta) = \sum_{i=1}^{I} I_i(\theta)$$

Para el caso del modelo 3PL, la función de información del ítem se calcula como

$$I_i(\theta) = a_i^2 \left[\frac{p_i - c_i}{1 - c_i}\right]^2 \left[\frac{1 - p_i}{p_i}\right]$$

Donde $P(U_{ij}=1 \mid \theta_j, a_i, b_i, c_i) = c_i + (1-c_i) \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}$

De $I_i(\theta)$ se puede observar que ítems con mayor discriminación a_i alcanzan una mayor función de información. Además, se puede demostrar que la función de información es mayor cuando el parámetro de pseudo-azar es $c_i=0$, que cuando toma un valor mayor a cero.

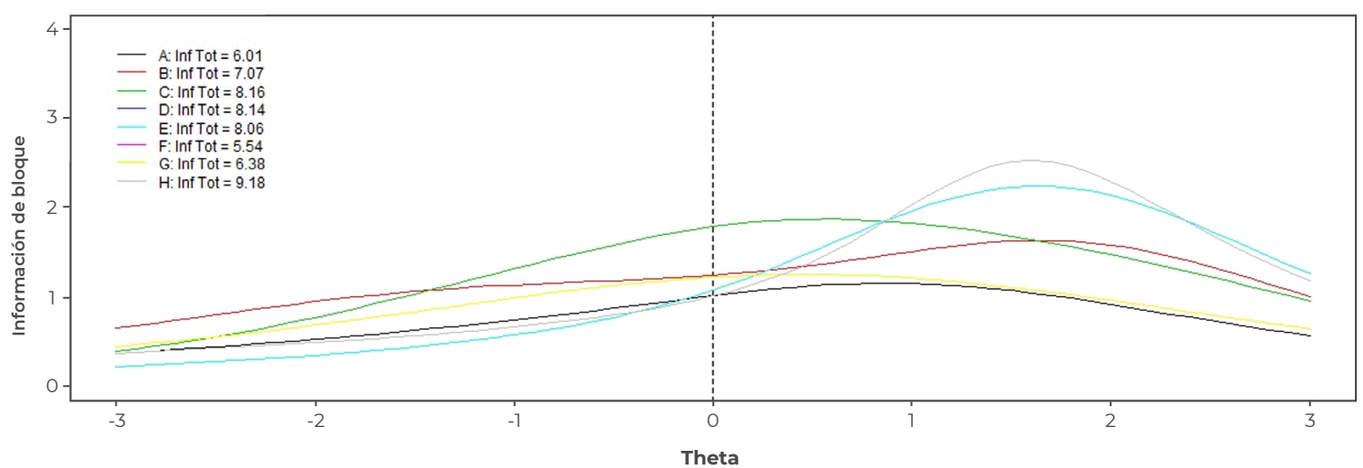
Para cumplir el objetivo de determinar la manera más adecuada de disminuir la longitud de la prueba Saber 11, utilizaremos la FIT. Dado que dicha función es la suma de las funciones de los ítems que componen la prueba, al remover ítems se disminuye la FIT y se aumenta el error de estimación. En este ejercicio se busca cómo seleccionar el conjunto de ítems tal que, al ser removido, se disminuya lo menos posible la FIT, de modo que se pueda reducir la longitud de la prueba y se impacte en la menor medida posible el error de medición de las habilidades θ.

3.4. Escenarios de eliminación de bloques

Dado que cada forma está compuesta por cuatro bloques en el armado de cada prueba, la estrategia consistió en calcular la función de información de cada bloque, de manera que aquel con la menor información sería eliminado de cada forma para la aplicación durante la emergencia sanitaria. Así, cada estudiante sería evaluado a partir de tres bloques por prueba en lugar de cuatro. El cálculo de las funciones de información de cada bloque se hizo con base en la metodología anteriormente presentada y utilizando la lista de ítems evaluados en el periodo 2018-2 con su respectivo armado, ya que las pruebas a aplicar en 2020-2 serían similares.

Para determinar el bloque con la menor información, se calcula la integral de la función de información $I(\theta)$ con base en los ítems que componen cada bloque. Como ejemplo, en la **Figura 4** se muestra la función de información para los bloques de la prueba de Lectura Crítica en 2018-2. El valor de la integral de la función de información se presenta en la esquina superior izquierda. A partir de este valor de la integral de la función de información para cada bloque, se puede determinar cuál de los cuatro bloques en cada forma se puede eliminar a fin de reducir la longitud de la prueba, de manera que se tenga el menor impacto en la precisión de la estimación de la habilidad de los evaluados en la versión corta de la prueba durante la emergencia sanitaria.

Figura 4. Curvas de información para los ocho bloques de la prueba de Lectura Crítica aplicados en 2018-2.



Nota: En la esquina superior izquierda se presenta para cada bloque el valor de la integral de su función de información.

Como se puede observar, algunos bloques tienen una función de información alta, como los bloques E y H, pero se concentran en las habilidades superiores (cercanas a 2 en escala logit). Esto indica que la información o precisión es alta para medir habilidades superiores, sin embargo, al mirar la curva de información alrededor de cero, la función es relativamente baja. Por lo tanto, esos bloques tienen una precisión relativamente baja para habilidades alrededor de cero, que es el promedio, en donde se concentran las habilidades de gran parte de la población.

Dado que algunos bloques pueden tener curvas de información altas, pero no necesariamente para las habilidades donde se encuentra la mayor parte de la población, se consideraron tres posibilidades para seleccionar los bloques a eliminar. Además de considerar la eliminación de bloques con menor información, se planteó como segunda opción remover los bloques con mayor sesgo (curvas desplazadas hacia la derecha), y, como tercera opción, eliminar los bloques a partir de algunos criterios escogidos por la Subdirección de Diseño de Instrumentos.

A partir de estas tres opciones para seleccionar los bloques que se eliminarían de la aplicación para 2020-2, se definieron algunos escenarios para simular cómo sería el error estándar en la calificación, con base en los datos de la prueba aplicada en 2018-2. Los escenarios se presentan en la **Figura 5,** en donde se consideran además algunas opciones para eliminar los ítems en la prueba de Inglés, la cual está conformada por partes en lugar de bloques. En la siguiente Sección se aclaran estos detalles para la prueba de Inglés, la cual tiene un diseño distinto a las demás pruebas de Saber 11

Figura 5. Escenarios considerados para seleccionar los bloques o partes de las pruebas que se eliminarían para reducir la longitud del cuadernillo a aplicar en 2020-2.

Escenario 1	Escenario 2	Escenario 3	Escenario 4	Escenario 5
Eliminación del bloque de menor información y del bloque de mayor sesgo (corrimiento a la derecha) para las pruebas de Lectura Crítica, Ciencias Naturales, Ciencias Sociales y Matemáticas	Eliminación de bloques según criterios de la SDI para las pruebas de Lectura Crítica, Ciencias Naturales, Ciencias Sociales y Matemáticas.	Eliminación del bloque de menor información y del bloque de mayor sesgo (corrimiento a la derecha) para las pruebas de Lectura Crítica, Ciencias Naturales, Ciencias Sociales y Matemáticas	Eliminación de bloques según criterios menor información para las pruebas de Lectura Crítica, Ciencias Naturales, Ciencias Sociales y Matemáticas.	Eliminación de bloques según criterios menor información para las pruebas de Lectura Crítica, Ciencias Naturales, Ciencias Sociales y Matemáticas.
Eliminación de partes de menor información (partes 3 y 6) para la prueba de inglés	Eliminación de partes según criterios de la SDI (partes 1 y 6) para la prueba de inglés.	Eliminación de partes según criterios de la SDI (partes 1 y 6) para la prueba de inglés.	Eliminación de partes de menor información (partes 3 y 6) para la prueba de inglés	Eliminación de partes de menor información (partes 1 y 6) para la prueba de inglés

Nota: En la parte inferior se describe la opción para la eliminación en la prueba de Inglés y en la parte superior para las demás pruebas.

Las simulaciones consistieron en tomar los datos de las pruebas de 2018-2 y eliminar los bloques o partes de acuerdo con los escenarios planteados, de modo que se acortaron virtualmente las pruebas que aplicaron los estudiantes en 2018-2. Al calificar las pruebas acortadas bajo estos escenarios, se calcularon los errores estándar de la estimación de las habilidades, y, con base en los resultados, se seleccionó la alternativa más adecuada para reducir la longitud de la prueba que presentarían los estudiantes en Saber 11 de 2020-2.

3.5. Escenarios de eliminación de partes

Como se mencionó anteriormente, la prueba de Inglés está compuesta por partes y no por formas como el resto de las pruebas. Esta prueba está conformada por 45 ítems agrupados en 7 partes y cada parte se centra en medir un nivel específico de manejo del inglés, de acuerdo con el marco común europeo de referencia (MCER). En la **Tabla 2**, se encuentra cada una de las partes del armado de la prueba, el nivel de inglés del MCER que evalúa esa parte, una clasificación del nivel de acuerdo con los parámetros del Icfes y el número de ítems. Se observan diferencias grandes en la cantidad de ítems entre algunas partes.

Para la prueba de Inglés se decide excluir 2 partes que permitan reducir la extensión de la prueba. Sin embargo, por la particularidad de esta prueba no se usan los criterios expuestos para la eliminación de bloques, sino tres criterios relacionados con las características, estructura de la prueba y distribución de evaluados por nivel. El primer criterio tiene en cuenta la cantidad de ítems de la parte excluida con el fin de controlar la longitud del recorte y no excluir una gran cantidad de ítems que reduzca demasiado la prueba. El segundo criterio se relaciona con el nivel de inglés que mide la parte con el fin de que no se excluyan dos partes del mismo nivel. El tercer criterio está relacionado con la cantidad de estudiantes que se clasifican en los niveles de inglés, buscando no remover las partes en los niveles donde existe gran cantidad de estudiantes evaluados.

Tabla 2. Estructura de la prueba de inglés en Saber 11°

Partes	Nivel	Clasificación de nivel	No. de ítems
Parte.1	A1	Principiante	5
Parte.2	A1	Principiante	5
Parte.3	A2	Básico	5
Parte.4	A2 y B1	Básico - Pre Intermedio	8
Parte.5	A2 y B1	Básico - Pre Intermedio	7
Parte.6	B1 y B2	Pre Intermedio - Intermedio	5
Parte.7	B1 y B2	Pre Intermedio - Intermedio	10
Total			**45**

En la **Tabla 3** se observa que un gran porcentaje de estudiantes se encuentran en los niveles de Inglés de menor desempeño. Aproximadamente el 90 % se encuentran en los niveles A- y A1 y A2, principalmente el nivel A-, seguido de A1 y en menor medida A2. El restante de evaluados que son alrededor del 10% se encuentran en niveles B1 y B+.

Los tres criterios para la eliminación de partes fueron tenidos en cuenta y analizados en conjunto entre la Subdirección de Diseño de Instrumentos y la Subdirección de Estadísticas, llegando a la propuesta de generar dos escenarios de eliminación. En el primero se eliminan las partes 1 y 6, mientras que en el segundo se eliminan las partes 3 y 6. En ambos casos se eliminan partes que se centran en diferentes niveles de inglés y partes compuestas por 5 ítems para que la reducción no fuera mayor a 10 ítems. En los dos escenarios se elimina la parte 6, ya que existen menos evaluados en niveles B1 y B2 con el fin de no aumentar el error de estimación en las otras partes que evalúan niveles más bajos de Inglés, donde se encuentra una mayor cantidad de estudiantes.

Tabla 3. Distribución de los estudiantes por nivel de inglés en las aplicaciones calendario A 2017-2019.

Periodo Cal A / Niveles*	A-	A1	A2	B1	B+	Población
2017-2	44%	30%	16%	8%	2%	464.871
2018-2	37%	34%	19%	8%	2%	463.030
2019-4	45%	30%	16%	7%	2%	475.312

3.6. Equiparación

La equiparación es un proceso estadístico que permite que las puntuaciones de una prueba sean comparables cuando esta es aplicada en diferentes momentos o cuando se utilizan diferentes formas para una prueba. El interés por realizar comparaciones de puntajes, en particular en el tiempo, radica en el uso que se le puede dar a esta información. A nivel individual, por ejemplo, los evaluados pueden hacerle seguimiento a su nivel de habilidad en aquello que mide la prueba y, a nivel institucional, la información recopilada sirve como insumo en la formulación y seguimiento de políticas públicas (Kolen y Brennan, 2014). En la aplicación de Saber 11 en 2020-2 es de particular interés el asegurar que la prueba sea comparable, por lo cual el proceso de equiparación es especialmente relevante.

Existen varias metodologías para realizar el proceso de equiparación cuando se utiliza un modelo de TRI para la calificación. En algunas propuestas se buscan constantes de equiparación (momentos de los parámetros, Stocking-Lord) que permiten hacer comparables los puntajes, mientras que en otras se realiza la equiparación en el proceso de estimación de parámetros de los ítems (calibración).

La metodología usada en la equiparación de las pruebas Saber 11 corresponde a la Fijación de Parámetros de Calibración del Ítem (FIPC, por sus siglas en inglés), la cual permite equiparar los puntajes de un nuevo periodo a la escala base. En esta metodología se fijan o anclan los parámetros de los ítems a la escala base (histórica) y se estiman los parámetros de los ítems nuevos en la aplicación, de tal forma que estos, y los puntajes de los evaluados, queden en la escala base (Kang y Petersen, 2012). Esta es la metodología empleada en todas las aplicaciones de Saber 11 desde 2014-2, y se emplea también en la versión corta de 2020-2.

Previo a la equiparación de una nueva aplicación, es necesario analizar si los ítems aplicados en la prueba tienen las mismas propiedades psicométricas en términos de sus parámetros (a_i, b_i y c_i) comparados a la escala histórica de las aplicaciones anteriores. Al tener ítems con los mismos parámetros en la nueva aplicación, se puede equiparar a través de esos ítems y se asegura que la calificación se encuentre en una misma escala. El análisis de Funcionamiento Diferencial del Ítem (DIF, por sus siglas en inglés) permite determinar cuáles ítems se comportan distinto entre aplicaciones y cuáles ítems se comportan de manera similar.

Por lo anterior, para realizar la implementación del FIPC, es necesario determinar si el comportamiento psicométrico de los ítems comunes entre aplicaciones es invariante o si, por el contrario, presenta un funcionamiento diferencial entre las diferentes aplicaciones. Para ello, se realiza un análisis de DIF sobre los ítems comunes en el cual se identifican los ítems sin DIF para los cuales se pueden fijar los parámetros. Los ítems que presentan DIF se recalibran en el proceso de calificación, es decir, se estima de nuevo sus parámetros con la población de la nueva aplicación en el análisis.

Debido a los cambios en la aplicación y en particular al recorte de la prueba Saber 11° en el año 2020 por la pandemia, es de particular interés comparar el comportamiento psicométrico de los ítems entre la prueba corta de 2020-2 y la prueba larga aplicada en periodos anteriores. Para garantizar la comparabilidad de los puntajes en las aplicaciones, es importante que haya un número suficiente de ítems sin DIF, de modo que los puntajes en las aplicaciones se puedan equiparar correctamente. Usualmente se requiere un número mínimo de 12 ítems sin DIF, o, en pruebas largas, al menos el 25% del número total de ítems en la prueba. Por lo tanto, a continuación, se explica de forma detallada la metodología de DIF que se utilizó para estos análisis.

3.7. Análisis de DIF

En general, se considera que un ítem presenta DIF cuando evaluados pertenecientes a distintos grupos y con un mismo nivel de habilidad cuentan con distintas probabilidades de responder correctamente el ítem; es decir, cuando las curvas características del ítem (CCI) difieren entre grupos, lo cual está relacionado con un sesgo en la medición. En el caso de la aplicación de la prueba Saber 11 en 2020-2, para determinar la invarianza en el tiempo, los grupos corresponden a: 1) la población de la calibración histórica relacionada con la evaluación en formato largo, y 2) la población de la aplicación 2020-2 en formato corto.

Teniendo en cuenta que con la metodología de FIPC se busca que los resultados de distintas aplicaciones de la prueba sean comparables, el análisis de DIF implementado en este caso se enfoca en determinar si el ítem presenta un funcionamiento diferencial entre la última aplicación (prueba corta) y las aplicaciones de años anteriores. A partir de esta identificación, los ítems sin DIF serán fijados o anclados a la calificación histórica, mientras que ítems con DIF deben ser objeto de un proceso adicional de recalibración para incluirse en la medición.

La metodología usada en los análisis de DIF se basa en la propuesta desarrollada por Raju (1988), que consiste en estudiar las diferencias de las CCI para los dos grupos de comparación; es decir, cuantificar la magnitud de las diferencias a lo largo del rasgo latente para identificar cuando esta es grande y, por ende, significa que el ítem presenta DIF (Ver **Figura 6**).

Dependiendo de la forma de las diferencias de la CCI, el funcionamiento diferencial se clasifica en uniforme o no uniforme. En el DIF uniforme, la probabilidad de contestar correctamente es uniformemente mayor en todo el rasgo latente respecto al otro grupo, mientras que en el DIF no uniforme la diferencia en la probabilidad de la respuesta correcta varía a lo largo del rasgo latente y las dos curvas se interceptan (Ver **Figura 7**).

Considerando los tamaños poblacionales en la evaluación y la estructura de armado de Bloques Incompletos Balanceados, la metodología implementada para hacer el análisis de DIF está basada en la propuesta de Oshima, Raju y Nanda (2006), la cual contempla la estimación del índice de DIF no compensatorio (NcDIF). Este estadístico se basa en la comparación de las CCI de dos grupos: un grupo base (focal) y un segundo grupo de interés (referencia).

Figura 6. Ejemplo de la diferencia de CCI de acuerdo con el enfoque de Raju.

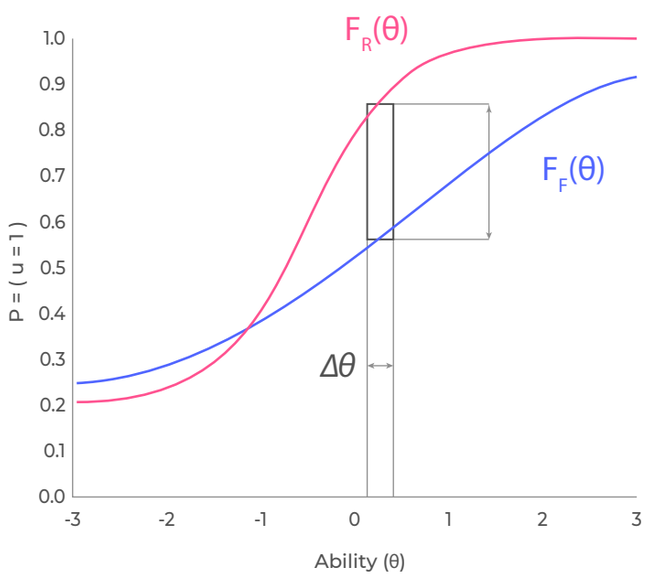
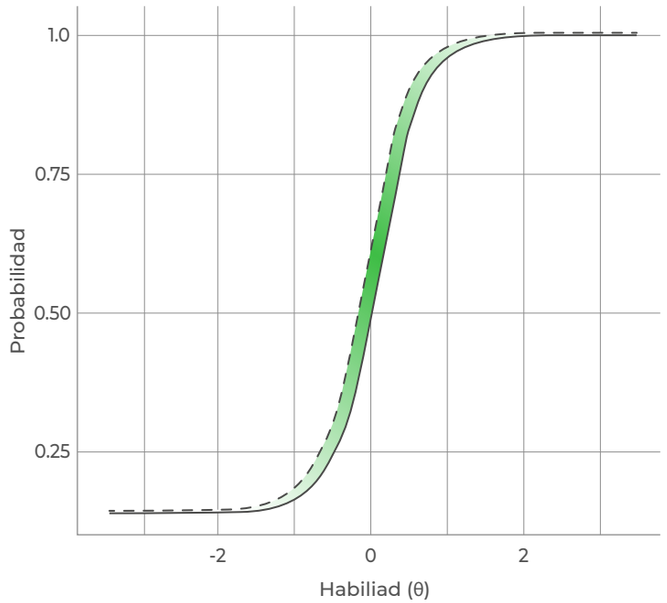
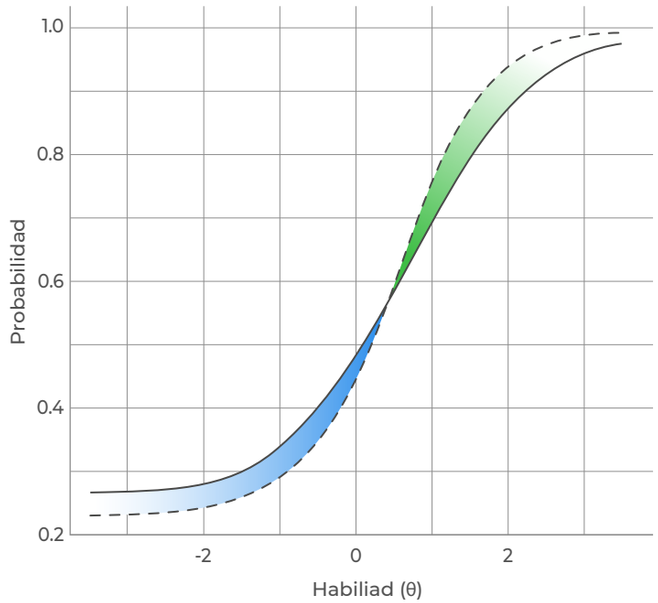


Figura 7. Ejemplos de la representación de DIF. Cada curva corresponde a cada grupo de comparación

a) DIF uniforme



b) DIF no uniforme



Para cada ítem dicotómico i, la brecha d_i entre los dos grupos de comparación es la diferencia en la probabilidad de una respuesta correcta entre ambos grupos en la habilidad θ de acuerdo con el modelo de calificación; es decir, para el evaluado j con habilidad θ_j la brecha está dada por:

$$d_i(\theta_j) = P_F(\theta_j) - P_R(\theta_j) = P_F(U_{ij}=1 \mid \theta=\theta_j) - P_R(U_{ij}=1 \mid \theta=\theta_j),$$

donde $P_F(\theta)$ y $P_F(\theta)$ son respectivamente la probabilidad de que un evaluado con habilidad θ responda correctamente el ítem i de acuerdo con las calibraciones para el grupo Focal y el grupo de Referencia, respectivamente. El índice NcDIF se calcula como el valor esperado sobre la distribución de la habilidad del grupo focal del cuadrado de esta brecha, esto es:

$$NcDIF_i = E_F(d_i^2(\theta)) = \int_{-\infty}^{\infty} (P_F(\theta) - P_R(\theta))^2 f_F(\theta) d\theta$$

donde la función f_F es la función de densidad de probabilidad de la habilidad del grupo focal. Este índice recibe el nombre de no compensatorio porque, al tomar el cuadrado de la brecha, las diferencias en direcciones opuestas no se cancelan entre sí.

Es pertinente tener en cuenta que, si bien el índice NcDIF es el valor esperado de la distancia entre las CCI del grupo base y el grupo de interés, en la integral se toma como referencia la distribución del primero al momento de calcular el índice. Por esta razón, en este contexto, la metodología permite determinar si la CCI en la aplicación más reciente puede considerarse igual a su versión histórica, o si en cambio presenta un cambio en el tiempo.

Para la implementación de la metodología de DIF, se debe realizar una serie de pasos. El primero es obtener las calibraciones de los ítems en las poblaciones de interés en los análisis; es decir, del grupo focal y del grupo referencia. En este caso, el grupo focal corresponde a las calibraciones históricas, y el grupo de referencia es el grupo de estudiantes que presentaron la prueba más corta en 2020-2. Luego, se colocan las calibraciones en una misma escala llevando las del grupo de referencia al grupo focal a través de un método de equiparación, como, por ejemplo, el método de Stocking-Lord (Kolen y Brennan, 2014). En el tercer paso, se cuantifica la diferencia de las curvas características de los ítems a través del índice NcDIF. Por último, se categoriza la magnitud de esta diferencia a través de la clasificación de la estadística utilizando un análisis de Tamaño del Efecto (TE). La metodología del TE utilizada en este ejercicio corresponde a la propuesta de Wright y Oshima (2015), en la cual se proponen tres categorías para evaluar el tamaño del DIF, que guardan relación con la propuesta del Delta de la Educational Testing Service (ETS): A, un efecto insignificante; B, un efecto moderado; y C, un efecto grande. Los puntos de corte para clasificar el TE del DIF de los ítems se encuentran en la **Tabla 4**.

Tabla 4. Valores de TE clasificando el índice NcDIF.

Tamaño del efecto	Clasificación
A	Despreciable
B	Moderado
C	Grande

3.8. Calificación de la prueba

La prueba se calificó de la misma forma como se califica el examen Saber 11 en cada periodo; es decir, se anclaron los ítems sin DIF a la escala histórica para asegurar que los puntajes de los estudiantes estuvieran en la misma escala de la línea base. En la equiparación se anclaron los ítems que no presentaron DIF y se liberaron los ítems que se ubicaron en la categoría C, para que sus parámetros sean reestimados bajo las nuevas condiciones de aplicación.

Como un paso final de análisis, se procedió a comparar los puntajes obtenidos por los estudiantes en esta versión corta de la prueba con los puntajes de la prueba original aplicada en los periodos anteriores. La comparación se realizó a partir de los puntajes promedio y desviaciones estándar.

Examen Saber 11°

04.

Resultados

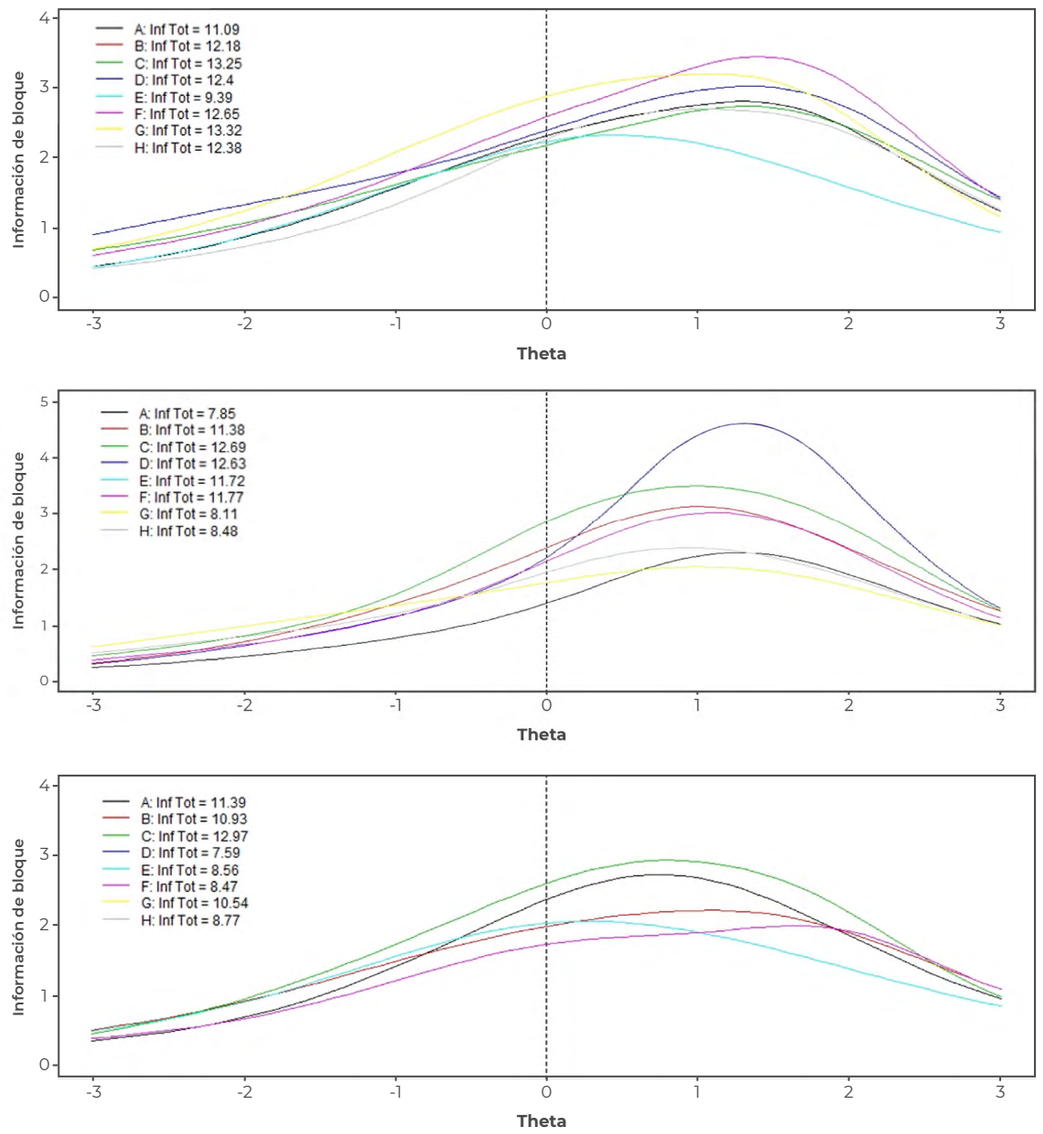
4.1. Curvas de información por bloque

Como se explicó anteriormente, se calculó la curva de información para cada uno de los bloques de las pruebas aplicadas en 2018-2. Las curvas de los bloques para Lectura Crítica se presentaron en la **Figura 5** y las curvas de información para Ciencias Naturales, Matemáticas y Ciencias Sociales se reportan en la **Figura 8**. Como se puede observar, los bloques brindan distintos niveles de información a las pruebas. Por ejemplo, en Ciencias Naturales, la integral de la curva para el bloque G es de 13,32, mientras que para el bloque E es igual a 9,39, lo cual se ve reflejado en que la curva del bloque G está por encima de la del bloque E a lo largo de las habilidades.

Si bien los bloques tienen una dificultad promedio igual a cero por construcción, las curvas de información tienden a ser sesgadas como se puede observar. Las curvas, en su mayoría, alcanzan su máximo en valores altos de la habilidad, lo cual quiere decir que los bloques de las pruebas tienden a tener ítems con dificultades altas. Esto se debe a que no es sencillo construir ítems con dificultades bajas. Como consecuencia, se tiene una mayor precisión para evaluar estudiantes en rangos de habilidad altos que en rangos bajos y medios.

Nota: En la esquina superior izquierda se presenta para cada bloque el valor de la integral de su función de información.

Figura 8. Curvas de información para los ocho bloques de tres pruebas aplicadas en 2018-2



a. *Curvas de información para la prueba de ciencias naturales.*

b. *Curvas de información para la prueba de matemáticas.*

c. *Curvas de información para la prueba de ciencias sociales*

4.2. Evaluación de escenarios - Efecto de eliminación del bloque y por partes

Con base en los escenarios planteados en la Figura 6, se llevaron a cabo las simulaciones eliminando los bloques o partes de la prueba que respondieron los estudiantes en 2018-2. La Tabla 6 reporta el puntaje promedio obtenido por los estudiantes a partir de la prueba con todos los bloques (Oficial) y el puntaje promedio al acortar las pruebas bajo los cinco escenarios (Escenario). Como se puede observar, el puntaje promedio con base en toda la prueba es igual al puntaje promedio bajo los cinco escenarios para Inglés y Lectura Crítica, mientras que para las otras tres pruebas se encuentran diferencias de un punto en algunos de los escenarios.

Por otro lado, se calculó el error estándar promedio de la estimación de las habilidades con la prueba completa y al recortar la prueba bajo los cinco escenarios. El error estándar se calcula a partir de la función de información de la prueba que presenta cada estudiante, con la metodología explicada anteriormente. Como se puede observar en la **Tabla 5**, el aumento en el error promedio al quitar uno de los bloques está entre 0,02 y 0,05 dependiendo de la prueba y del escenario. En la última columna se reporta el aumento porcentual en el error estándar promedio.

Tabla 5. Valores de TE clasificando el índice NcDIF.

Prueba	Escenario	Puntaje promedio		Error promedio		Diferencia porcentual del error
		Oficial	Escenario	Oficial	Escenario	
Ciencias Naturales	1	50	50	0,33	0,35	7,12%
	2	50	50	0,33	0,35	7,52%
	3	50	50	0,33	0,35	7,12%
	4	50	49	0,33	0,34	5,58%
	5	50	49	0,33	0,34	5,58%
Ciencias Sociales	1	48	49	0,34	0,38	10,27%
	2	48	48	0,34	0,37	7,64%
	3	48	49	0,34	0,38	10,27%
	4	48	49	0,34	0,37	8,54%
	5	48	49	0,34	0,37	8,54%
Inglés	1	51	51	0,36	0,38	4,26%
	2	51	51	0,36	0,4	9,88%
	3	51	51	0,36	0,4	9,65%
	4	51	51	0,37	0,39	4,96%
	5	51	51	0,37	0,4	9,85%

Continua en la siguiente página

Para todas las pruebas, el escenario 4 conlleva al menor aumento en el error estándar promedio, o el segundo menor aumento, excepto para Matemáticas. Por lo tanto, esta fue la alternativa que se recomendó para reducir la longitud de la prueba que presentaron los estudiantes en Saber 11 de 2020-2. Es decir, que los estudiantes presentaron tres bloques en vez de cuatro, y el bloque a remover sería aquel que aportaba menor información a la prueba. A continuación, se presentan el análisis de DIF para comparar el comportamiento de los ítems en la prueba corta y en las aplicaciones anteriores. Por último, se discuten los resultados con respecto a los puntajes de los estudiantes que presentaron esta versión más corta del examen.

Tabla 5. Valores de TE clasificando el índice NcDIF.

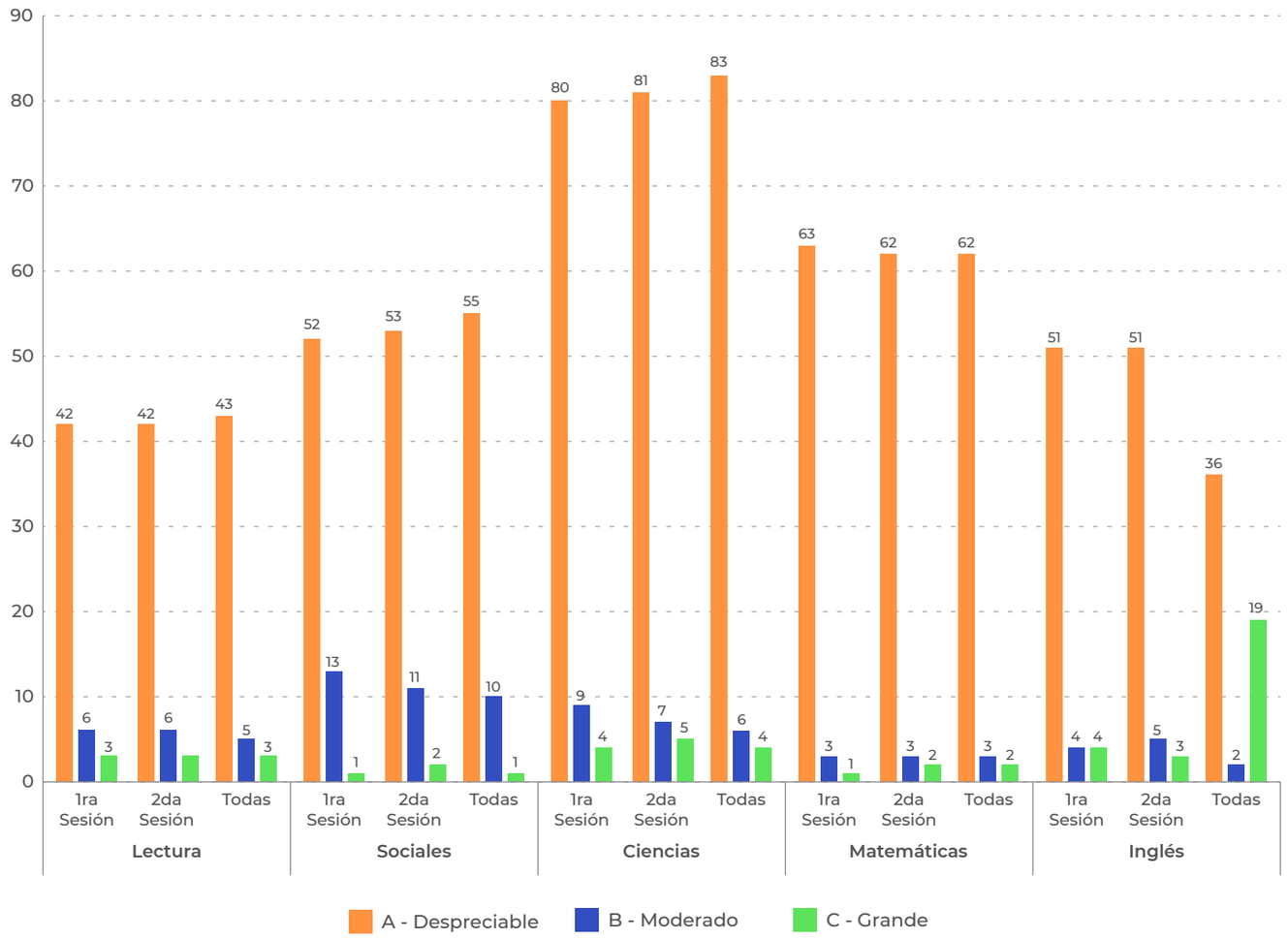
Prueba	Escenario	Puntaje promedio		Error promedio		Diferencia porcentual del error
		Oficial	Escenario	Oficial	Escenario	
Lectura Crítica	1	53	53	0,41	0,45	8,46%
	2	53	53	0,42	0,46	10,20%
	3	53	53	0,41	0,45	8,46%
	4	53	53	0,42	0,46	9,94%
	5	53	53	0,42	0,46	9,94%
Matemáticas	1	50	50	0,34	0,38	9,62%
	2	50	51	0,35	0,4	13,81%
	3	50	50	0,34	0,38	9,62%
	4	50	50	0,35	0,39	11,70%
		50	51	0,35	0,39	12,31%

4.3. Análisis de DIF

Como ocurre con cada nueva aplicación, se realiza análisis de DIF para determinar si hay cambios importantes en el comportamiento de algunos de los ítems. Esto toma aún más importancia en la aplicación de Saber 11 de 2020-2 dados los cambios realizados en la logística de aplicación, e incluso, en el diseño de armado. Por lo anterior, se realizaron tres tipos de análisis de DIF teniendo en cuenta la sesión en la que los estudiantes presentaron la prueba: 1) análisis de DIF para los ítems calibrados de los evaluados en la primera sesión de la mañana contra los resultados históricos, 2) análisis de DIF para los ítems calibrados de los evaluados en la segunda aplicación en la tarde contra los resultados históricos y 3) análisis de DIF para los ítems calibrados con toda la población (primera y segunda sesión) contra los resultados históricos. Teniendo en cuenta lo anterior, la Figura 9 presenta los resultados obtenidos en los análisis al comparar cada una de las poblaciones con las calibraciones históricas.

En cuanto a la comparación entre sesiones, se visualiza un comportamiento uniforme entre las dos sesiones. Cuando se hace uso de toda la población, se observa un aumento considerable en la cantidad de ítems con DIF ubicados en la categoría C para la prueba de Inglés, respecto a lo proyectado individualmente en cada una de las dos sesiones. Este comportamiento no se observa para las otras pruebas. En general, se observa que hay un buen número de ítems que no se encuentran en categoría C para todas las pruebas, de manera que se cuenta con la cantidad necesaria de ítems para equiparar con las aplicaciones anteriores y asegurar que los puntajes se encuentren en la misma escala y sean comparables.

Figura 9. Comparación de los resultados del análisis de DIF por sesión de aplicación y con toda la población para el examen Saber 11 – calendario A 2020.



4.4. Calificación de la prueba en formato largo y corto

Después de aplicar la prueba corta y posterior al proceso de calificación, se analiza el comportamiento de los puntajes obtenidos equiparando la aplicación de 2020-2 a la escala histórica. Teniendo en cuenta que normalmente la prueba Saber 11 se presenta en formato largo y en el 2020-2 se presentó en formato corto, a continuación, se realizan comparaciones con base en los promedios y desviaciones estándar para analizar si hay cambios importantes en el comportamiento de los puntajes. En la Tabla 6 se reportan las estadísticas para los años 2018, 2019 y 2020 para los estudiantes de calendario A. Allí se observa que, entre el 2018 y 2019, el promedio del puntaje disminuyó levemente en las pruebas de Sociales y ciudadanas, Ciencias naturales, Inglés y Lectura crítica; mientras que para la prueba de matemáticas aumentó levemente. Sin embargo, ente el 2019 y 2020, los promedios del puntaje de estas pruebas no variaron respecto al periodo anterior con el que se comparó, exceptuando la de Inglés, que disminuyó levemente. Respecto al promedio del puntaje global, este ha disminuido 4,62 puntos entre el 2018 y 2019, mientras que este promedio aumentó 2,18 puntos entre el 2019 y 2020.

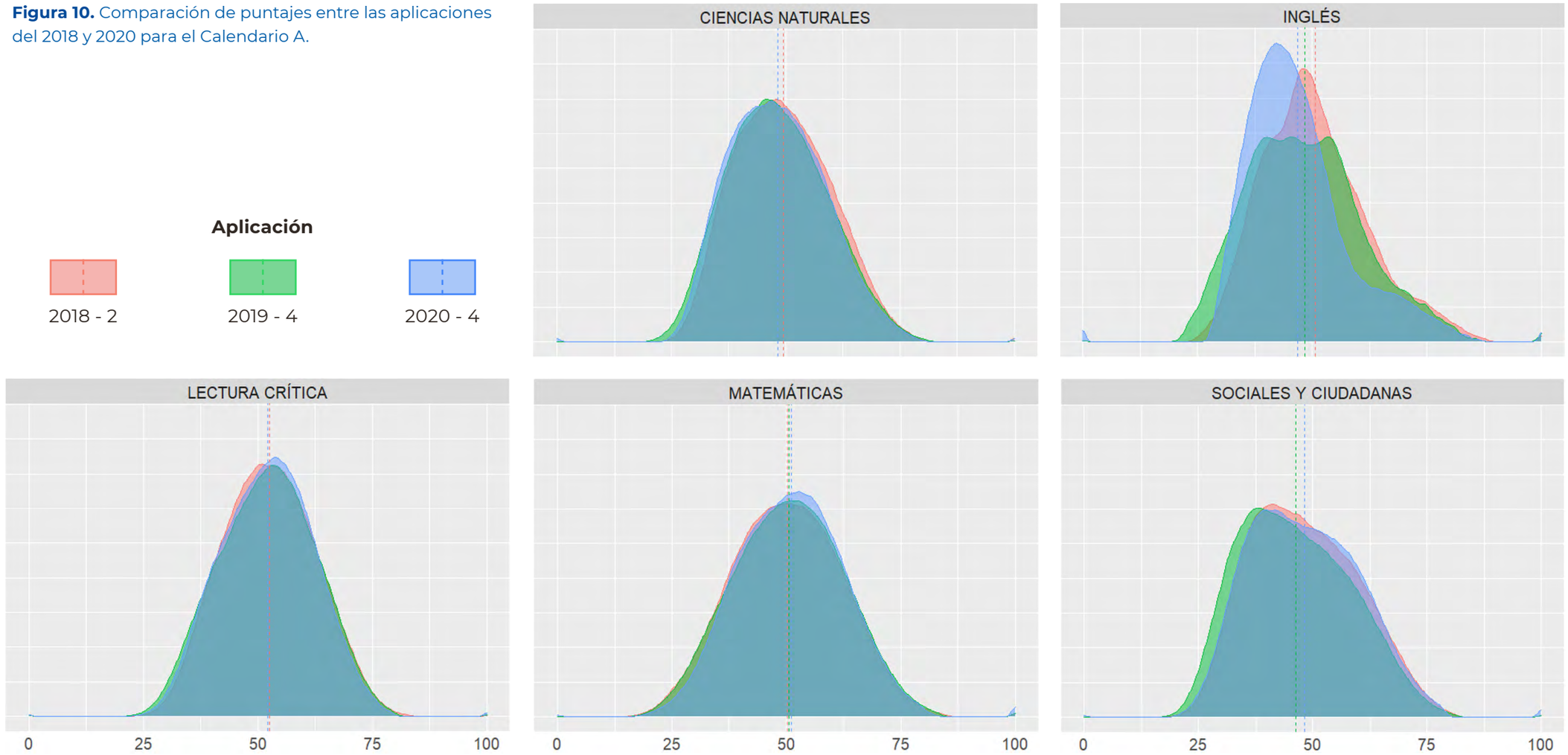
Respecto a la desviación estándar, se identifican variaciones leves entre las aplicaciones comparadas. Sin embargo, para el puntaje global se identifica que la desviación disminuyó 2,69 puntos entre el 2019 y 2020.

Por otra parte, la **Figura 10** permite visualizar la distribución de los puntajes para los tres periodos comparados. Como se puede observar, las distribuciones son similares y no se observan cambios importantes entre las aplicaciones. Sin embargo, en la prueba de Inglés, la distribución de la aplicación del año 2019 tiende a ser más platicúrtica respecto a las otras aplicaciones; mientras que para el año 2020 la distribución se concentra levemente hacia la izquierda y toma una forma más leptocúrtica respecto a los años anteriores. Esto está relacionado con el hecho que inglés presentó mayor disminución del promedio en los años comparados.

Tabla 6. Promedio del puntaje y desviación estándar para los módulos generales, según el periodo de aplicación

Prueba	Promedio del puntaje			Desviación		
	2018	2019	2020	2018	2019	2020
Ciencias Naturales	50	48	48	10	11	10
Inglés	51	48	47	12	13	11
Lectura Crítica	53	52	52	10	11	10
Matemáticas	50	51	51	12	12	12
Sociales y ciudadanas	48	46	48	12	12	12
Puntaje global	251	246	248	50	51	49

Figura 10. Comparación de puntajes entre las aplicaciones del 2018 y 2020 para el Calendario A.



Conclusiones

Debido a la pandemia, fue necesario realizar ajustes en la prueba Saber 11 para poder llevar a cabo su aplicación durante el 2020. Para la primera aplicación a los estudiantes de calendario B, bastó con conseguir más colegios, de manera que se tuvieran menos estudiantes por salón y se pudiera llevar a cabo la prueba con distanciamiento social. Para calendario A, fue necesario reducir la longitud de la prueba para que los estudiantes pudieran resolver el examen en una sesión en lugar de dos. De esta manera, se pudo evaluar la mitad de los estudiantes en la mañana y la otra mitad en la tarde, con lo cual se pudo asegurar también el distanciamiento social.

Para disminuir la longitud de la prueba en calendario A, se tomaron dos medidas: remover los ítems piloto de los cuadernillos y evaluar cada estudiante con cuadernillos de tres bloques en vez de cuatro para cada prueba. Para seleccionar cuál sería el bloque para remover en la prueba, se hizo uso de la función de información, de manera que se eliminó el bloque que aportaba menos información, y que, por lo tanto, tenía un menor impacto en la precisión de la estimación de la habilidad de los evaluados.

El proceso de equiparación usado para 2020-2 fue el mismo que se utiliza regularmente, que es, a través de fijación de parámetros. Dado que el análisis de DIF indicó que había un número alto de ítems sin comportamiento diferencial, fue posible equiparar los puntajes de la prueba de 2020-2 con la escala histórica, de manera que los puntajes fueran comparables. Al analizar de manera descriptiva los puntajes de 2020-2, se encontró que no hubo un cambio importante en comparación a aplicaciones anteriores, lo cual corrobora que el acortamiento de la prueba no afectó la validez de los resultados.

Es importante recalcar que, aunque los puntajes de esta versión reducida de la prueba son comparables con la escala histórica, no es recomendable mantener la versión corta para futuras aplicaciones, ya que esto aumenta de alguna manera el error estándar de la estimación de las habilidades de los estudiantes. Por lo tanto, es mejor tener la mayor precisión posible para asegurar las mejores condiciones en la evaluación realizada por el Instituto. Además, se deben volver a incluir los ítems piloto en aplicaciones posteriores para poder continuar aumentando el banco de ítems disponibles para las pruebas.

Referencias

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

De Ayala, R. J. (2018). Item response theory and Rasch modeling. The reviewer's guide to quantitative methods in the social sciences, 145-163.

Instituto Colombiano para la Evaluación de la Educación (4 de septiembre de 2019). Resolución 675/2019. Por la cual se reglamenta el proceso de inscripción a los exámenes que realiza el Icfes.

Instituto Colombiano para la Evaluación de la Educación (2020). Informe Nacional de resultados Saber 11 2020, Bogotá, Colombia: autor.

Instituto Colombiano para la Evaluación de la Educación (2021). Guía de orientación Saber 11. 2021-1, Icfes, Bogotá, Colombia: autor.

Kang, T. y Petersen, N. S. (2012). Linking item parameters to a base scale. Asia Pacific Education Review, 13, 311–321.

Kolen, M. J., y Brennan, R. L. (2014). Test equating, scaling, and linking: Methods and practices (3.a ed.). Berlín, Alemania: Springer Science + Business Media. DOI: 10.1007/978-1-4939-0317-7

Oshima, T., Raju, N. y Nanda, A. (2006). A New Method for Assessing the Statistical Significance in the Differential Functioning of Items and Tests (DFIT) Framework. Journal of Educational Measurement, 43(1). 1 -17. DOI: https://doi.org/10.1111/j.1745-3984.2006.00001.

Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika, 53(4), 495–502. doi:10.1007/bf02294403

Wright, K. D., y Oshima, T. C. (2015). An Effect Size Measure for Raju's Differential Functioning for Items and Tests. Educational and psychological measurement, 75(2), 338–358. DOI: https://doi.org/10.1177/0013164414532944

Calle 26 N.° 69-76, Torre 2, Piso 15, Edificio Elemento, Bogotá, D. C., Colombia
www.icfes.gov.co