MINISTERIO DE EDUCACIÓN NACIONAL

Informe técnico sobre los examenes de Estado

en tiempos de la COVID-19

Saber TyT y Saber Pro

Subdirección de Estadísticas
Dirección de Evaluación

Bogotá, diciembre de 2022

Examen Saber TyT | Examen Saber Pro | icfes

MINISTERIO DE EDUCACIÓN NACIONAL

Elaboración del documento
John Alexander Calderón Rodríguez
Luis Adrián Quintero Sarmiento
Karen Rosana Córdoba Perozo
Nelson Andrés Rodríguez Rivera
Andrés Ricardo Rodríguez Nagles
Nila Fernanda Amaya Melo

DIseño y diagramación
Kevin Ostos Peñaloza

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

ISBN: 978-958-11-1012-4

Bogotá D.C., diciembre 2022

Informe técnico sobre los examenes de Estado en tiempos de la COVID-19

Saber TyT y Saber Pro

Examen Saber TyT

Examen Saber Pro

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 TyT y Saber Pro se aplica de manera periódica, principalmente a los estudiantes que están a punto de finalizar la educación superior en Colombia. Sin embargo, ante la emergencia sanitaria asociada a la COVID-19, el Instituto requirió pasar de aplicar la prueba en papel y aplicarla de manera electrónica, esto con el fin de permitir que los estudiantes realizaran el examen desde sus casas. Posteriormente, se habilitaron centros con computadores para que los evaluados sin los recursos tecnológicos en sus casas pudieran llevar a cabo la prueba. Dado este cambio, resultó necesario verificar que los resultados de la prueba aplicada en computador fueran comparables con los resultados de la prueba aplicada anteriormente en papel. Esto es importante para garantizar la comparabilidad de los puntajes en el tiempo como lo indica la Ley 1324 de 2009.

En esta línea, el presente documento parte desde la importancia de la aplicación de los exámenes, posteriormente describe algunos estudios relevantes en la literatura sobre la comparación de pruebas en papel y en computador, la metodología utilizada en el presente estudio para comparar los formatos de aplicación, y, finalmente, presenta las principales conclusiones sobre el cambio de formato realizado durante la pandemia.

.01

Sobre el examen
Saber TyT y Pro

El Instituto Colombiano para la Evaluación de la Educación (Icfes) tiene como misión evaluar el cumplimiento de objetivos planteados para el sector educativo a través de los exámenes de estado para los distintos niveles de educación (artículo 1, Ley 1324 de 2009). Dentro de los instrumentos contemplados para los fines mencionados se encuentran los exámenes Saber Pro y TyT, los cuales buscan corroborar el nivel en el que se encuentran las competencias de los estudiantes y recoger información que le permita a las Instituciones de Educación Superior (IES) tomar decisiones para mejorar la calidad de sus programas.

Las pruebas Saber TyT y Pro son presentadas por estudiantes de carreras técnicas y profesionales que hayan aprobado el 75% del programa académico en curso o graduados que desean presentar el examen de forma independiente (art. 4, Decreto 3963 de 2009). Por año se presentan aproximadamente 160.000 evaluados para el examen Saber TyT, y alrededor de 260.000 personas para el examen Saber Pro.

Antes de la pandemia, los exámenes estaban compuestos por módulos genéricos y módulos específicos. Estos módulos son evaluados en dos sesiones. En la primera sesión se presentan los módulos genéricos, que se componen de 5 pruebas que evalúan las competencias de Lectura crítica, Razonamiento cuantitativo, Competencias ciudadanas, Comunicación escrita e Inglés. La aplicación de esta primera parte del examen tiene una duración de 4 horas y 20 minutos. En la segunda sesión se aplican los módulos específicos, esta sesión tiene una duración de 1 hora, en el caso que presenten un solo módulo; sin embargo, si el evaluado presenta 2 módulos, este tiempo cambia a 2 horas máximo.

Los módulos específicos están dirigidos únicamente a estudiantes que presentan el examen por primera vez y que son inscritos por las IES a las que pertenecen. Para el examen Saber Pro, estos módulos se componen de 1 a 3 pruebas seleccionadas por las IES en donde se evalúan temáticas específicas que se relacionan con el programa al que pertenecen las personas evaluadas. En el examen Saber TyT los evaluados presentan un solo módulo específico que también es definido por las IES en las que se encuentran.

Los exámenes Saber Pro y TyT contaron con un formato de papel y lápiz hasta el 2019. Debido a la emergencia de la COVID-19, la aplicación de estas dos pruebas cambió a formato electrónico y se permitió a los evaluados presentar la prueba de manera remota, teniendo en cuenta las recomendaciones de prevención y distanciamiento físico establecidas por el Gobierno Nacional (Icfes, 2020a; Icfes 2020b).

Para implementar estas pruebas en formato electrónico durante la pandemia, se aprovecharon las experiencias previas del instituto, las cuales se discuten a continuación.

1.1. Antecedentes

1.1.1. Pruebas en formatos electrónicos

Con el fin de innovar y fortalecer los procesos de evaluación que desarrolla el Icfes, en los últimos años se han desarrollado pruebas en formatos electrónicos, las cuales tienen múltiples ventajas tales como reducir los costos por impresión y transporte del material, además de permitir que los estudiantes tomen la prueba en distintos momentos en centros de cómputo o desde su casa.

La herramienta creada por el Instituto para la aplicación de exámenes en formato electrónico se denomina PLEXI (PLataforma de presentación de EXámenes del Icfes). Esta es una aplicación que se instala en el computador donde el estudiante presenta la prueba. Dentro de las herramientas de PLEXI se incluye un cronómetro para que los evaluados tengan control del tiempo durante el examen. Además, se ofrece la opción de resaltar textos, una lupa para aumentar el tamaño de las imágenes, así como la posibilidad de que el estudiante pueda navegar a lo largo de la prueba y devolverse a preguntas que ya se hayan resuelto previamente (Icfes, 2020c).

A través de PLEXI, el Icfes ha desarrollado diversas pruebas en formato electrónico. De acuerdo con el Informe de Gestión de 2019 entregado por el Icfes, durante ese año se realizaron 15 pruebas electrónicas por medio de PLEXI (Ver **Tabla 1**).

En el caso de la prueba Saber 11 INSOR, la plataforma permitió que los evaluados contaran con un video por cada ítem para brindar acompañamiento de intérprete. La aplicación de esta prueba en el calendario B se realizó en 5 departamentos y contó con 9 evaluados, mientras que la aplicación para el calendario A se realizó en 26 departamentos y contó con 387 evaluados.

Frente a Avancemos 4°, 6° y 8°, esta prueba tuvo dos aplicaciones de manera electrónica en el 2019 y se logró evaluar 2.909 estudiantes en la primera y 484.957 en la segunda. En este caso, PLEXI permitió que los resultados se le entregaran a los evaluados en un corto periodo de tiempo.

Respecto a las pruebas Saber TyT 2019 - Policía Nacional extemporánea, estas se aplicaron también por medio de PLEXI, utilizando los equipos de cómputo que se encontraran en los sitios de aplicación. El total de evaluados por medio del aplicativo fue de 4.397.

Por otra parte, PLEXI fue la herramienta que permitió la integración de la metodología de pruebas adaptativas por computador (CAT, por sus siglas en inglés), que consiste en asignar ítems a los evaluados de manera progresiva de acuerdo con su desempeño en las preguntas anteriores (Icfes, 2019b). Esta metodología brinda mayor flexibilidad y requiere un menor número de preguntas para evaluar a los estudiantes con una precisión adecuada. La implementación de la metodología CAT se realizó con el acompañamiento técnico del profesor Mark Reckase de la Universidad de Michigan.

El avance en la integración de la metodología CAT al Icfes dio como resultado El Icfes tiene un preicfes, que es una herramienta virtual a la que pueden acceder gratuitamente los estudiantes que desean familiarizarse con las preguntas de los exámenes de Estado.

Prueba	No. Evaluados
Saber TyT 2019 extemporánea – Policía Nacional	4.397
Saber 11° INSOR Calendario B	9
Avancemos 4°, 6° y 8° - Primer semestre	292.305
Pre Saber Adaptativo	1.406
Pre Saber Electrónico	1.518
ECDF – Selección de pares evaluadores	3.380
Piloto PISA	255
Saber 11° INSOR Calendario A	352
Avancemos 4°, 6° y 8° - Primer semestre	250.613
Prueba de ascenso para mayores (Policía Nacional)	50
Saber Pro y Saber TyT en el exterior (Comunicación escrita)	1.882
Avancemos 4°, 6° y 8° - Edición Chocó	547
Saber 3°, 5° y 9° - Prueba piloto	5.894
Saber Pro y Saber TyT extemporánea	317
TOTAL	**567360**

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

Nota: Tomado de Icfes (2019a).

Por otra parte, PLEXI fue la herramienta que permitió la integración de la metodología de pruebas adaptativas por computador (CAT, por sus siglas en inglés), que consiste en asignar ítems a los evaluados de manera progresiva de acuerdo con su desempeño en las preguntas anteriores (Icfes, 2019b). Esta metodología brinda mayor flexibilidad y requiere un menor número de preguntas para evaluar a los estudiantes con una precisión adecuada. La implementación de la metodología CAT se realizó con el acompañamiento técnico del profesor Mark Reckase de la Universidad de Michigan.

El avance en la integración de la metodología CAT al Icfes dio como resultado El Icfes tiene un preicfes, que es una herramienta virtual a la que pueden acceder gratuitamente los estudiantes que desean familiarizarse con las preguntas de los exámenes de Estado.

1.1.2. Pruebas en formato papel vs. formato electrónico

Cuando se desea aplicar una prueba simultáneamente tanto en lápiz y papel como en computador, es importante determinar si los puntajes obtenidos a partir de los dos vehículos de aplicación son comparables. El Icfes ha desarrollado varios estudios para analizar dicha comparabilidad. Por ejemplo, el piloto de Saber 3, 5, 9 en 2019 fue aplicado en papel, pero se tuvo una submuestra de colegios que contaban con equipos de cómputo y conexión a internet, de manera que se les aplicó la prueba de manera electrónica. En Quintero et al. (2022), se analizó la comparabilidad de la prueba Saber 3, 5, 9 en los formatos de papel y computador, utilizando métodos cuasiexperimentales y se encontró que hay diferencias significativas para grado 9 únicamente. Sin embargo, dado que la muestra de colegios que aplicaron la prueba electrónica no fue seleccionada de manera aleatoria, sino que se seleccionó a conveniencia de acuerdo con la disponibilidad de computadores e internet, se debe tener cuidado con el alcance de las conclusiones.

Además, se encontró que el número de ítems con comportamiento diferencial entre los dos formatos (DIF, por sus siglas en inglés), disminuye a medida que aumenta el grado de escolaridad. Al revisar el texto y la presentación de los ítems en la plataforma, se concluyó que un elemento determinante para la presencia de DIF es que los estudiantes tengan que desplazarse en la plataforma de la prueba electrónica para leer toda la pregunta y las opciones de respuesta (scrolling).

Por otra parte, en el 2021, el Icfes realizó un estudio en conjunto con el Ministerio de Educación de la República Dominicana para recolectar evidencias sobre las semejanzas y diferencias entre las Pruebas Nacionales aplicadas en papel y en computador a estudiantes de sexto de secundaria de la República Dominicana (Icfes, 2022). El estudio arrojó diferencias significativas entre los promedios en los dos formatos para las pruebas de Matemática y Lengua Española. El tamaño del efecto resultó pequeño para Matemáticas, con puntajes más altos para los estudiantes que presentaron la prueba en computador, y mediano para Lengua Española, a favor de la aplicación en papel.

Estos estudios de comparación de formatos realizados por el Icfes, se han llevado a cabo con estudiantes de educación básica y media. En otros estudios realizados se ha encontrado que a mayor escolaridad de los evaluados, hay mayor comparabilidad entre los dos formatos (Choi & Tinkler, 2002). Por lo tanto, en el presente documento se verificará la comparabilidad para estudiantes de educación superior.

1.2. Implicaciones COVID-19

Teniendo en cuenta las restricciones de aislamiento asociadas a la COVID-19, el Icfes tuvo que replantear algunos aspectos logísticos en la aplicación de las pruebas para garantizar el cumplimiento de las medidas de bioseguridad. Específicamente, para las pruebas Saber TyT y Saber Pro, la implementación del formato electrónico fue una propuesta ante la situación de la pandemia para no generar contacto entre estudiantes. La primera aplicación en este formato se realizó con la prueba Saber TyT 2020-1, en donde se permitió a los evaluados realizarla desde casa, si los estudiantes contaban con los elementos tecnológicos para presentar el examen. En las aplicaciones de Saber TyT 2020-3 y Saber Pro 2020-3, la prueba continuó en formato electrónico y

esta podía ser presentada en casa o en sitio, si no tenían acceso a equipos. En sitio se disponían los equipos para que los evaluados asistieran durante las sesiones de aplicación definidas por el Icfes. El análisis en el presente estudio se enfocará en las pruebas Saber TyT 2020-1 y 2020-3 y Saber Pro 2020-3. Las pruebas Saber TyT 2020-2, Saber Pro 2020-1 y Saber Pro 2020-2 corresponden a aplicaciones pequeñas en el exterior.

Otro ajuste que se realizó en estas pruebas durante la pandemia fue disminuir el tiempo de aplicación, por lo que no se aplicaron los módulos de competencias específicas y la estructura del examen contó solo con las preguntas de los módulos genéricos. Por otro lado, no se incluyeron ítems de pilotaje en estas aplicaciones.

Además, se generaron varias sesiones de aplicación en diferentes días y jornadas debido a que la cantidad de personas que presentaron el examen Saber TyT Y Saber Pro fue alta. Esto asegura el correcto funcionamiento de la plataforma, ya que no es recomendable tener un número muy alto de evaluados aplicando la prueba al tiempo.

1.3. Información general frente a las aplicaciones realizadas durante la pandemia

Los datos en el presente estudio corresponden a los estudiantes que se inscribieron y presentaron las pruebas Saber TyT 2020-1, Saber TyT 2020-3 y Saber Pro 2020-3. El número de evaluados en cada opción (sitio, casa) se presenta en la **Tabla 2**. Dada la situación de la pandemia en el primer semestre de 2020, la aplicación de las pruebas Saber TyT 2020-1 se desarrolló sólo en las casas de los evaluados. Para Saber TyT 2020-3 se tuvo alrededor del 50% de los estudiantes en casa y en sitio, mientras que para Saber Pro la mayoría de los estudiantes (82%) realizaron el examen desde su casa. Esto se debe en parte a que, para aplicar la prueba en casa, se debe disponer de un computador y de conexión a internet, lo cual es más común en los estudiantes que presentan el examen Saber Pro.

Durante la aplicación, los estudiantes respondieron un cuestionario socioeconómico junto con las pruebas de competencias básicas. En dicho cuestionario se recogió información acerca de la educación y ocupación de los padres de los evaluados, posesiones en el hogar y valor de la matrícula, entre otros. Además, se cuenta con información de las IES en donde se encuentran matriculados los estudiantes, como es su zona, sector, y puntajes promedio en aplicaciones anteriores de las pruebas Saber Pro y Saber TyT. Esta información es útil para caracterizar los datos y analizar su comparabilidad con los evaluados que presentaban las pruebas antes de la pandemia.

Tabla 2. Número de evaluados en casa y en sitio para las pruebas Saber TyT y Saber Pro 2020.

Examen	Casa	Sitio
Saber TyT 2020-1	80.907	No aplica[1]
Saber TyT 2020-3	42.194	41.255
Saber Pro 2020-3	212.069	45.821

1 Para la aplicación de 2020-1 de las pruebas Saber TyT, únicamente se realizó aplicación en casa, debido a la emergencia sanitaria que se presentó en ese momento.

1.4. Monitoreo de estudiantes

Teniendo en cuenta que la aplicación fue virtual, el monitoreo de los evaluados se realizó en diferentes etapas. Primero, se aplicó un proceso de verificación a partir del reconocimiento facial con el fin de comprobar que la persona que está en el computador es la que se inscribió y luego de esta verificación, se iniciaba con la presentación del examen.

Durante la emergencia sanitaria, se cambió la definición de conductas prohibidas para la presentación de las pruebas de estado por medio de la Resolución 530 de 2020, la cual modifica el artículo 4 de la Resolución 631 de 2015. Dado lo anterior, las conductas prohibidas que se establecieron para la aplicación electrónica fueron las siguientes:

- Interactuar o hablar con otras personas.
- Consumir bebidas alcohólicas o sustancias psicoactivas, así como presentar el examen bajo los efectos de alguna de estas sustancias.
- Manipular libros, cuadernos, hojas con o sin anotaciones, exceptuando el espacio durante el cual se está presentado el módulo de razonamiento cuantitativo, ya que en ese caso se podía usar una hoja y un lápiz para hacer las operaciones necesarias.
- Usar calculadoras, celulares, cámaras fotográficas, reproductores musicales, gafas inteligentes, reloj inteligente, ni ningún otro tipo de aparato no autorizado, solo se pueden utilizar en caso de que estén autorizadas para las personas con discapacidad.
- Capturar fotos de la pantalla del computador.
- Ausentarse de la cámara web durante la aplicación. Esto se consideró como fraude en caso que el software de vigilancia capturara la imagen de supervisión sin que apareciera el examinando, en once (11) o más capturas sucesivas en la misma sesión; o veinte (20) o más capturas en toda la sesión.
- Acceder a otras páginas web o aplicaciones durante la presentación del examen.
- Usar gafas, cachuchas, tapabocas, diademas o cualquier elemento que impida la correcta visualización del rostro a través de la cámara web.

Además, el Icfes contó con un software de inteligencia artificial para auditar la aplicación de forma remota, por medio de un registró fotográfico que se tomó a los evaluados cada 15 segundos a través de la cámara web.

Así mismo, la plataforma fue monitoreada por personal humano para realizar seguimiento a las alertas que generó el software frente a los comportamientos de los evaluados. Si el evaluado no corregía la falta por la cual se está generando la alerta y el comportamiento se mantenía por 14 minutos aproximadamente, se anulaba el examen. Es importante mencionar que, durante la aplicación, se solicitó a los evaluados la validación del rostro en 2 ocasiones, al iniciar cada sesión. Dado lo anterior, las medidas tomadas para realizar el monitoreo requirieron que el Icfes fortaleciera su acompañamiento a los evaluados durante todo el proceso. (Icfes, 2020c).

02.

Metodología para la implementación de ajustes en los exámenes Saber TyT y Pro **debido a la emergencia sanitaria**

Dentro de los procesos que realiza el Icfes, el instituto debe garantizar la comparabilidad de los puntajes de las pruebas Saber Pro y Saber TyT durante al menos 12 años a partir de la línea base (2014-2) (parágrafo b, art. 7, Ley 1324 de 2009). Por lo tanto, al cambiar el formato de aplicación en papel y lápiz a una aplicación electrónica durante la pandemia, se generaron acciones para garantizar que los puntajes se puedan mantener en una misma escala de calificación, de manera que se asegure la comparabilidad de los puntajes previamente obtenidos en papel con los resultantes de la prueba electrónica. A continuación, se presenta la metodología utilizada para comparar los dos vehículos de aplicación.

2.1. Comparabilidad de formatos

De acuerdo con Berman et al. (2020) la comparabilidad implica que, idealmente, estudiantes con el mismo puntaje son igualmente competentes en el trazo latente que la prueba desea medir, sin importar el formato de aplicación. En otras palabras, un estudiante con cierta habilidad recibiría el mismo puntaje si realiza la prueba en cualquiera de los dos formatos.

Existen varias razones por las cuales pueden diferir los puntajes al aplicar la prueba en papel y en computador, ya que no es exactamente lo mismo responder una pregunta en los dos formatos, especialmente cuando el evaluado no está familiarizado con el uso de herramientas tecnológicas de este tipo. Un ejemplo común ocurre cuando las preguntas son largas y no se pueden leer completamente en la pantalla, sino que se debe hacer un desplazamiento (scrolling) para visualizar todo el texto. Tener gráficos y tablas también puede tener un impacto, teniendo en cuenta que estos elementos no se pueden manipular y rayar de la misma forma en papel que en el computador.

Un estudio relacionado con la comparabilidad de los dos vehículos de aplicación es el de Choi & Tinkler (2002), en donde se encontró que los efectos de formato son mayores en estudiantes de grados escolares inferiores, en comparación a los estudiantes de grados más altos. En este sentido, se podría esperar que las diferencias al cambiar el formato de aplicación en las pruebas Saber Pro y Saber TyT sean pequeñas, ya que los evaluados son de un nivel de escolaridad alto, y, además, se espera que tengan en general un buen grado de familiaridad con herramientas electrónicas. En todo caso, es necesario verificar si se puede asumir que los puntajes de los exámenes Saber Pro y Saber TyT son comparables al cambiar el formato de aplicación durante la emergencia sanitaria en el 2020.

En la literatura, la equivalencia entre los dos formatos se evalúa en dos niveles: 1) a nivel de las preguntas, mediante un análisis de funcionamiento diferencial del ítem (Bennett et al., 2008), y 2) de manera más global a nivel de la prueba, comparando los puntajes promedio obtenidos bajo los dos vehículos de aplicación (Hardcastle et al., 2017). A nivel de las preguntas, se puede determinar cuáles ítems tienen un funcionamiento similar sin importar el formato en el que el evaluado responde. En contraste, a nivel de la prueba, se analiza de manera más global si los estudiantes tienen puntajes similares en papel y en computador.

2.2. Comparación de poblaciones

El escenario ideal para analizar la comparabilidad de pruebas en papel y en computador es asignando una muestra de manera aleatoria a cada vehículo de aplicación. De esta manera, si se encuentran diferencias entre los formatos, se puede concluir que se deben únicamente al vehículo de aplicación. Sin embargo, es difícil tener muestras aleatorias para estudios de comparación entre papel y computador, ya que el llevar los dispositivos electrónicos y conectividad a los estudiantes en la muestra de computador resulta costoso. Por lo tanto, una alternativa común en estudios de comparación de formatos es el uso de cuasiexperimentos (Hardcastle et al., 2017). En dicha metodología se seleccionan muestras a conveniencia de acuerdo con la disponibilidad de equipos electrónicos en los colegios y conectividad. Para el análisis de resultados, se comparan las dos muestras de evaluados para determinar si hay diferencias visibles con respecto a las covariables disponibles, o si, por el contrario, no hay diferencias, y se puede asumir que las dos muestras son comparables.

Si las dos muestras son equivalentes, se puede realizar el análisis para comparar los resultados en papel y en computador. De esta manera, si se observan diferencias

en el desempeño en los dos formatos, se puede asumir que se deben al vehículo de aplicación, ya que no hay otras disimilitudes visibles entre las dos muestras que puedan explicar las diferencias en los resultados. Por otro lado, si no hay equivalencia entre las dos muestras de evaluados obtenidas a conveniencia, una alternativa común es aplicar métodos de pareo (Diamond & Sekhon, 2013) para remover esas discrepancias y disminuir posibles sesgos entre los resultados obtenidos por los estudiantes en los dos grupos. Una vez removidos los posibles sesgos, se procede a la comparación de resultados entre las muestras de papel y computador para obtener conclusiones sobre los vehículos de aplicación.

Para la comparación de las dos muestras, se puede analizar si hay diferencias significativas para las covariables disponibles en los datos. Sin embargo, para conjuntos de datos grandes, se pueden encontrar diferencias significativas debido al tamaño de la muestra, aun cuando las diferencias sean pequeñas. Por lo tanto, generalmente se utiliza el Tamaño del Efecto (TE), el cual es una diferencia estandarizada del promedio (o proporción) para cada covariable en el análisis. Para covariables numéricas, el TE se calcula de la siguiente manera:

$$TE = \frac{\mu_C - \mu_P}{\sigma_P}$$

Donde μ_C y μ_P son, respectivamente, el promedio de la covariable en el grupo de computador y el grupo de papel, mientras que σ_P es la desviación estándar de la muestra que tomó la prueba en papel. Para variables categóricas, se compara para cada categoría la proporción en las dos muestras de la siguiente forma:

$$TE = 2\,(arcsin\,\sqrt{p_C} - arcsin\,\sqrt{p_P})$$

Donde p_C y p_P son las proporciones en las muestras de computador y papel, respectivamente. La interpretación del TE se hace a partir de los puntos de corte presentados en la **Tabla 3**, de manera que la diferencia entre las dos muestras para cada covariable se puede clasificar como despreciable, pequeña, mediana o grande.

Tabla 3. Valores de TE clasificados por rangos.

Tamaño del efecto	Clasificación
TE < 0.2	Despreciable
0,2 ≤ TE ≤ 0.5	Pequeño
0,5 ≤ TE ≤ 0.8	Mediano
TE ≥ 0.8	Grande

Al comparar las dos muestras cuando no se seleccionaron de manera aleatoria, se espera que el TE sea pequeño o despreciable para las covariables disponibles. En caso de que se encuentren diferencias medianas o grandes, se deben utilizar métodos de emparejamiento como se mencionó anteriormente para tratar de hacer las muestras comparables.

2.3. Equiparación

La equiparación es un proceso estadístico que permite que las puntuaciones de una prueba sean comparables cuando esta es aplicada a diferentes poblaciones 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).

Existen varias metodologías para realizar el proceso de equiparación, especialmente cuando se utiliza un modelo de Teoría de Respuesta al Ítem (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 TyT y Pro 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).

Para calificar las pruebas de los estudiantes en una misma escala para los dos formatos, fue necesario determinar cuáles ítems se comportan de manera similar en formato computador (aplicaciones 2020) y papel (aplicaciones anteriores). Esto permite asumir que los parámetros de dichos ítems en el modelo de 3PL son los mismos en los dos formatos. Al tener ítems con los mismos parámetros en los dos formatos, se asegura que la calificación se encuentre en una misma escala y se pueden comparar de manera directa los puntajes obtenidos en papel y en computador. El análisis de Funcionamiento Diferencial del Ítem (DIF, por sus siglas en inglés) permite determinar cuáles ítems se comportan distinto en los dos formatos 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. Esta fase se realiza con el fin de establecer si es adecuado utilizar los parámetros históricos en el proceso de equiparación, o si, por el contrario, algunos ítems se comportan de manera distinta en la última aplicación con respecto a la escala histórica, y por lo tanto sería incorrecto asumir los mismos valores para sus parámetros. Para ello, se realiza un análisis de DIF sobre los ítems comunes entre aplicaciones, en el cual se identifican los ítems para los cuales se pueden fijar los parámetros y los ítems que presentan DIF con el fin de recalibrarlos, es decir, estimar de nuevo sus parámetros con la población de la aplicación en el análisis.

Debido a los cambios que se realizaron en la aplicación de las pruebas Saber TyT y Pro en 2020 por la pandemia, se debe comparar el comportamiento psicométrico de los ítems entre la prueba en computador y la prueba en papel aplicada en periodos anteriores. Para garantizar la comparabilidad de los puntajes en los dos formatos es importante que haya un número suficiente de ítems sin DIF, de modo que los puntajes en los dos formatos se puedan equiparar correctamente. Usualmente la literatura sugiere un número mínimo de 15 ítems sin DIF, o, en pruebas largas, al menos el 20% del número total de ítems en la prueba, criterio mencionado en varios artículos en particular en Kang y Petersen (2012).

2.4. 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 sesgo en la medición que, para este contexto, se refiere a favorecer a un formato de aplicación sobre otro en la evaluación. En el caso de las aplicaciones de las pruebas Saber TyT y Pro en el 2020, para determinar la invarianza en el tiempo, los grupos corresponden a: 1) la población de la calibración histórica relacionada con la evaluación en formato papel, y 2) la población de la aplicación 2020 en formato electrónico.

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 (electrónica) y las aplicaciones de años anteriores (histórica), conocida en la literatura como Item Parameter Drift (IPDRIFT). A partir de esta identificación los ítems sin DIF serán fijados o anclados en la calificación, 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 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 1**).

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 2**).

Considerando los tamaños poblacionales en la evaluación y la estructura de armado de Bloques Incompletos Balanceados, la metodología implementada para hacer el análisis de DRIFT 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 1. Ejemplo de la diferencia de CCI de acuerdo con el enfoque de Raju.

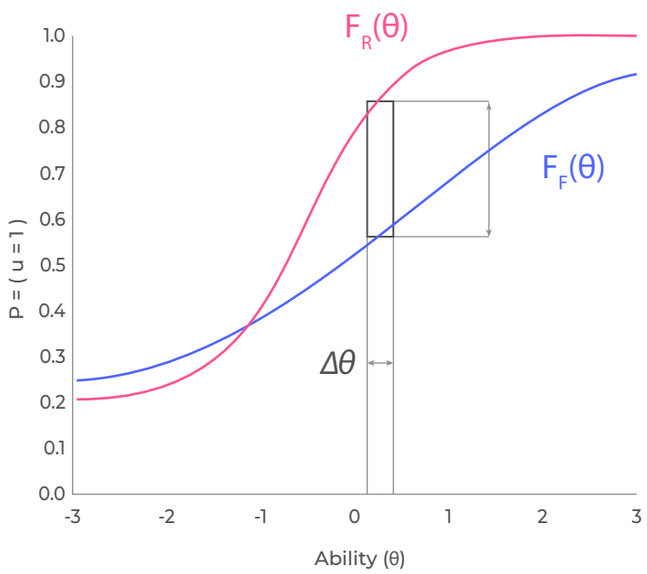
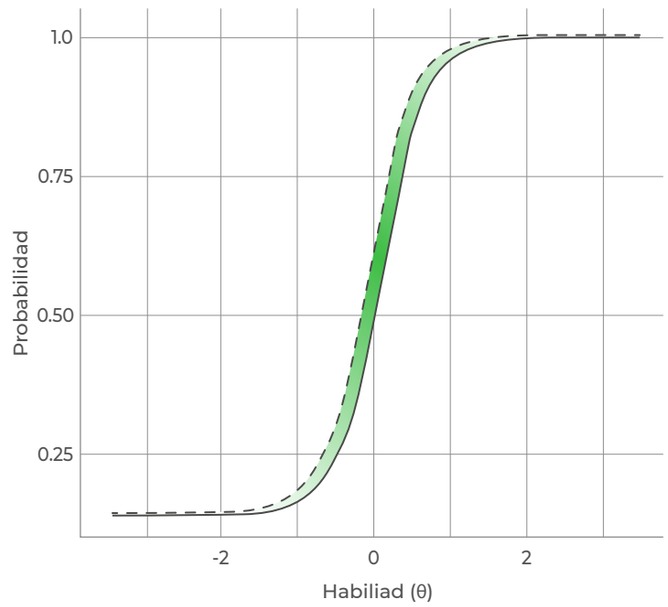
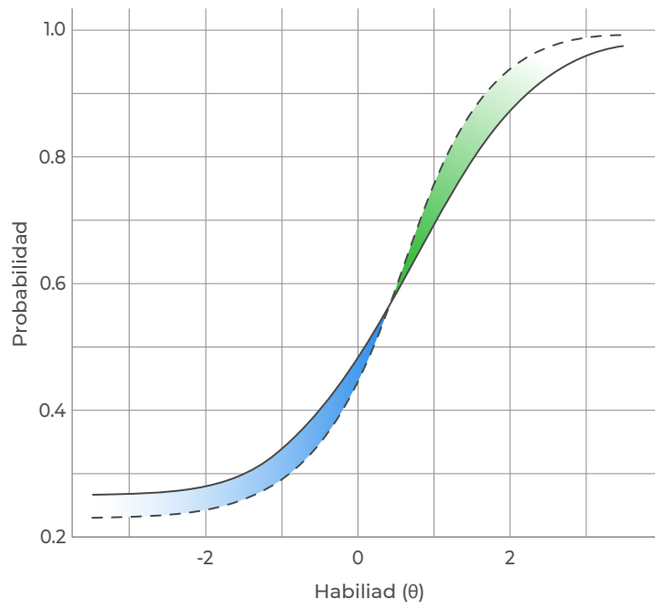


Figura 2. Ejemplos de la representación de DIF. Cada curva corresponde a cada grupo de comparación, en este caso, a cada formato

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 $P_R(\theta)$ 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 el contexto del IPDRIFT, 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 en papel, y el grupo de referencia es el grupo de estudiantes que presentaron la prueba en formato electrónico. 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 3.**

2.5. Otros análisis de DIF

Adicionalmente, con el fin de complementar los anteriores análisis, se decidió explorar una alternativa no paramétrica para el análisis de DIF que no dependiera de los supuestos del modelo TRI. Para este fin, se utilizaron los estadísticos Delta de Mantel-Haenzsel (MH) (Mantel y Haenzsel, 1959) y ΔR^2 de dos modelos de regresión logística anidados (Log) (Nagelkerke, 1991), siguiendo la clasificación del tamaño del efecto dada por Jodoin and Gierl (2001).

Ambos estadísticos, al igual que el índice NcDIF, permiten clasificar el tamaño del efecto de DIF de los ítems como despreciable, moderado o grande. Por otra parte, estos estadísticos presentan la ventaja de no depender de la estimación de parámetros de un modelo TRI sino únicamente de las respuestas dadas por los estudiantes a los ítems.

Para la metodología utilizada en este ejercicio, inicialmente se obtienen las clasificaciones del tamaño del efecto del DIF a partir de ambos estadísticos (MH y Log), y se concluye que un ítem presenta DIF si su tamaño del efecto es moderado o grande para cualquiera de los dos estadísticos.

2.6. Comparación de PRC de ítems

Teniendo en cuenta el cambio del formato de presentación de la prueba, el aumento de las sesiones de aplicación durante la pandemia y el lugar de aplicación de la prueba, se complementaron los análisis de ítem usuales en un proceso de calificación con un análisis adicional, con el fin de identificar ítems con comportamientos irregulares en las formas, sesiones o modalidad de aplicación de la prueba a través del Porcentaje de Respuestas Correctas (PRC) de los ítems. Estas irregularidades podrían ser consecuencia de la falta de tiempo durante la aplicación generada por retrasos en el ingreso del evaluado a la sesión, desconexión por problemas tecnológicos durante el examen u otras situaciones efecto de la dinámica de aplicación en el formato electrónico que podrían afectar el funcionamiento del ítem, en particular los ubicados al final de la prueba.

En este análisis para cada ítem se calcula el PRC con toda la población y para cada subpoblación definida por la sesión de aplicación y forma asignada. Teniendo en cuenta que en estas pruebas se presentan gran cantidad de estudiantes y por ende los tamaños poblacionales son grandes, se utiliza el TE para proporciones con el fin de comparar el comportamiento de los ítems a través de sus PRC entre sesiones, formas y modalidad de aplicación para identificar ítems irregulares.

El interés es identificar ítems con diferencias de PRC no despreciables que permitan identificar comportamientos irregulares (en algunas sesiones, por ejemplo), para así hacer un seguimiento detallado del ítem y determinar si sucedieron eventos atípicos durante la aplicación electrónica para algunos estudiantes. En análisis de los PRC se complementa con el análisis de ítem o el análisis de DIF, para así profundizar en la razón de posibles comportamientos atípicos.

2.7. Calificación de pruebas electrónicas

A partir de la información obtenida en los análisis mencionados, se tomó la decisión de calificar la prueba electrónica anclando los ítems sin DIF a la escala histórica. Esto se realizó con el fin de asegurar que los puntajes quedaran en la misma escala de la línea base y además, de liberar aquellos ítems que se ubicaron en la categoría C en alguno de los tres tipos de análisis, para que sus parámetros fueran reestimados bajo las nuevas condiciones de aplicación.

Teniendo los puntajes de la prueba electrónica en la escala base, se procedió a complementar el análisis de equivalencia entre formatos a partir de los puntajes promedio y desviaciones estándar de los 3 últimos periodos de aplicación, en donde el más reciente se había aplicado en computador, mientras que las dos aplicaciones anteriores, en lápiz y papel.

03.

Resultados

Examen Saber Pro

Examen Saber TyT

A continuación, se presentan los resultados obtenidos para la comparabilidad de formatos, teniendo en cuenta los análisis propuestos en el apartado de metodología. Estos resultados serán presentados según la aplicación: Saber TyT 2020-1, Saber TyT 2020-3 y Saber Pro 2020-3.

3.1. Análisis de resultados pruebas Saber TyT 2020-1

En este apartado, inicialmente, se presentan los resultados de comparar la población que presentó la prueba electrónica durante la pandemia con la población que presentaba la versión en papel anteriormente. Esto es útil para determinar si las poblaciones son comparables, de manera que se puedan realizar comparaciones directas entre los resultados en los dos formatos de aplicación, y verificar que, en caso de observar diferencias en los puntajes, éstas se puedan atribuir al cambio en el formato de aplicación, y no a un cambio en la población evaluada durante la pandemia.

Posteriormente, se realiza un análisis de DIF para determinar el número de ítems que tienen un comportamiento diferencial al ser aplicados en papel y en computador. De esta manera, se puede evaluar la posibilidad de tener una misma escala en el reporte histórico para los dos formatos, equiparando en caso de que haya un número adecuado de ítems sin DIF. Por último, se analiza si hay ítems con comportamientos irregulares con respecto al PRC obtenidos por los estudiantes en las sesiones y formas de la prueba.

3.1.1. Comparación de poblaciones

La prueba Saber TyT en 2020-1 se aplicó solamente en casa, dadas las restricciones de movilidad y de bioseguridad durante ese momento de la pandemia. La modalidad de aplicación en sitio se inició posteriormente tanto para Saber TyT como para Saber Pro. Por lo tanto, una condición necesaria para que los estudiantes pudieran aplicar Saber TyT en 2020-1 era que debían tener computador y conexión a internet en casa. Esta restricción genera la pregunta de si la población que presentó la prueba electrónica en esa aplicación podía ser distinta de la población regular que presentaba la prueba en las aplicaciones anteriores en papel. Por lo tanto, fue necesario determinar si esto ocurre, o si, por el contrario, se podía asumir que la población en 2020-1 era similar a las poblaciones que presentaban la prueba anteriormente, en cuyo caso se podía realizar análisis de DIF, entre otros, sin necesidad de utilizar métodos de pareo para tener poblaciones comparables en el análisis.

Para esto, se comparó el conjunto de estudiantes que presentaron Saber TyT en 2020-1 con la población que presentó Saber TyT en 2018-1; aplicación que se llevó a cabo en papel antes de la pandemia. Se tomó esta población ya que se esperaba que el armado de la prueba a aplicar de manera electrónica en 2020 fuera muy similar a la de 2018-1. La comparación se realizó a partir del tamaño del efecto (TE) utilizando las variables disponibles en el cuestionario sociodemográfico diligenciado por los estudiantes y usando algunas variables con las que se cuenta para las IES (zona, sector, puntajes promedio en 2019).

La **Tabla 4** presenta las variables de posesiones en el cuestionario sociodemográfico y el porcentaje de evaluados que tienen cada elemento. Además, para las variables relacionadas con el puntaje, se reporta el promedio de los puntajes de las IES a las que pertenecen los estudiantes en escala logit. Se presentan los valores para las aplicaciones 2018-1 y para 2020-1, así como la diferencia entre los dos periodos, y el TE para determinar qué tan grande es la diferencia. Como se puede observar, el TE es pequeño para las variables de internet y computador, y despreciable para las demás variables, tanto de tenencia, como de puntaje promedio de las IES a las que pertenecen los evaluados.

Estos resultados eran de esperarse, ya que para participar en la aplicación electrónica de Saber TyT 2020-1 era necesario que los estudiantes tuvieran computador y conexión a internet en casa, por lo cual, los evaluados difieren en estas características con respecto a la aplicación de 2018-1. Sin embargo, las dos

poblaciones resultan comparables con respecto a las otras características, ya que el TE es insignificante para las demás variables de posesiones y puntajes. El TE también resulta insignificante para las variables de educación de padres, ocupación de padres, estrato, cantidad de horas de trabajo laboral y región, las cuales no se reportan acá por la cantidad de subcategorías que tienen dichas variables y la longitud de la tabla que las incluye a todas, pero en todos estos casos, el TE estuvo por debajo de 0,2 en valor absoluto.

Con base en los resultados, se puede asumir que la población que presentó la prueba electrónica de Saber TyT en 2020-1 es comparable con las poblaciones que presentaban la prueba antes de pandemia, dado que solo se encuentran diferencias entre las dos muestras con respecto a tener computador y conexión a internet. Se puede considerar que los estudiantes que aplicaron Saber TyT en 2020-1 son similares para las demás variables, comparado con los evaluados que presentaron Saber TyT en 2018-1, la cual es la aplicación que se tomó como referencia de comparación para la población que presentaba la prueba en papel antes de la pandemia. Esto permite hacer los análisis en las siguientes secciones de manera más directa, ya que no es necesario utilizar métodos de pareo para crear subgrupos equivalentes previo al análisis de comparabilidad de formatos.

Tabla 4. Comparación de estudiantes que presentaron Saber TyT en 2018-1 (prueba en papel) y en 2020-1 (prueba en computador), la diferencia entre las dos aplicaciones y el tamaño del efecto (TE)

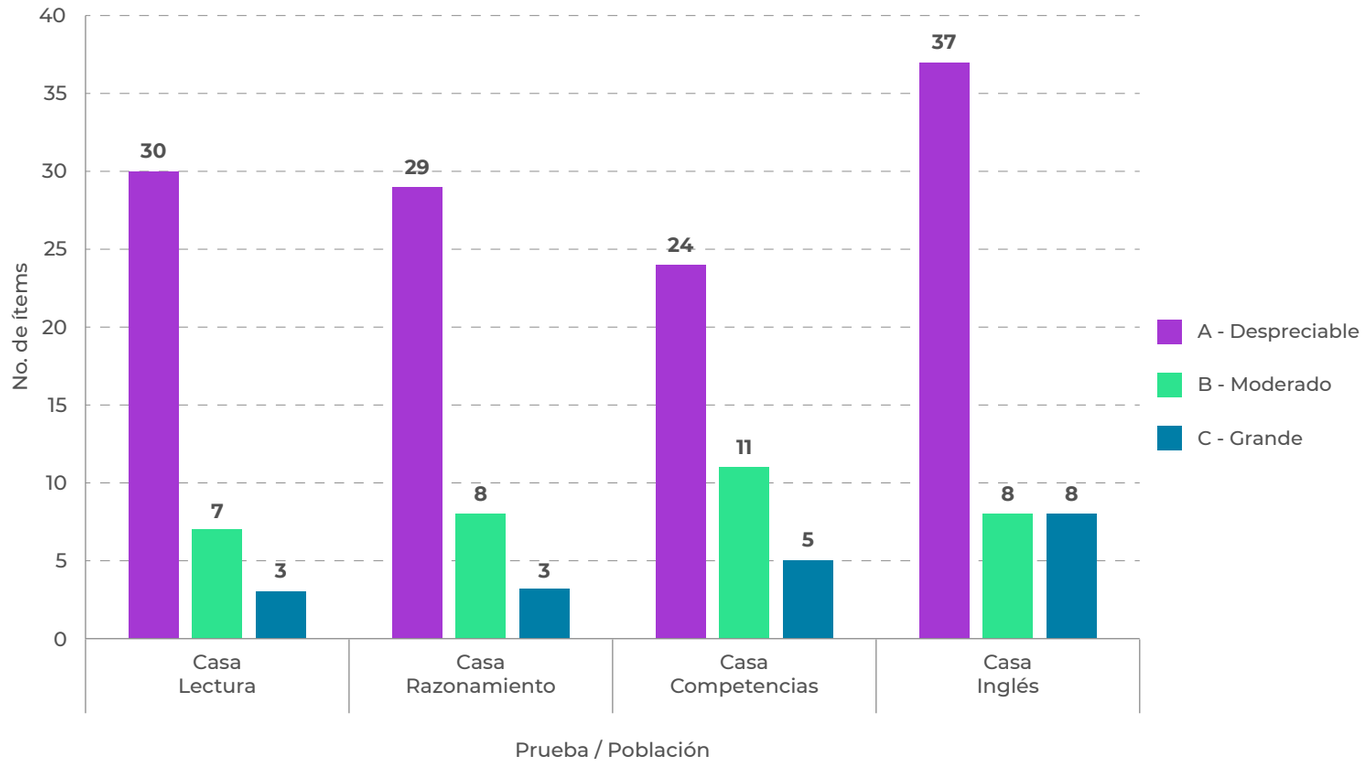
Variable	2018-1	2020-1	Diferencia	TE	Magnitud TE
Tiene internet	0,722	0,877	-0,155	-0,393	Pequeña
Tiene servicio cerrado de televisión	0,755	0,735	0,02	0,047	Despreciable
Tiene computador	0,765	0,871	-0,106	-0,276	Pequeña
Tiene lavadora	0,798	0,801	-0,003	-0,006	Despreciable
Tiene horno	0,429	0,399	0,03	0,06	Despreciable
Tiene automóvil particular	0,191	0,176	0,015	0,039	Despreciable
Tiene moto	0,395	0,376	0,019	0,039	Despreciable
Tiene consola de videojuegos	0,198	0,167	0,031	0,079	Despreciable
Puntaje Inglés 2019	-0,052	-0,083	0,031	-0,119	Despreciable
Puntaje Competencias Ciudadanas 2019	-0,538	-0,558	0,02	-0,085	Despreciable
Puntaje Lectura Crítica 2019	-0,263	-0,289	0,026	-0,113	Despreciable
Puntaje razonamiento Cuan. 2019	-0,618	-0,64	0,022	-0,09	Despreciable

3.1.2. Análisis de DIF

La estrategia para analizar si existían comportamientos diferenciales en los ítems a partir del formato de aplicación consistió en realizar un análisis de DIF entre papel y computador, como se explicó en la sección de metodología. De esta forma, se realizó análisis de DIF para los ítems calibrados con los evaluados en la aplicación Saber TyT 2020-1 que realizaron la prueba en Casa, contra los resultados históricos. Teniendo en cuenta lo anterior, a continuación, se presentan los resultados obtenidos en los análisis.

La **Figura 3** presenta la cantidad de ítems en cada categoría de magnitud de DIF de acuerdo con los análisis en cada una de las pruebas evaluadas. En general, se observa que la mayoría de los ítems se clasificaron de acuerdo con su estadística en la categoría A con DIF despreciable. En particular, la prueba de Competencias ciudadanas junto con la prueba de Inglés son las que presentan mayor número de ítems con DIF clasificados en las categorías moderado (B) y grande (C) en los análisis.

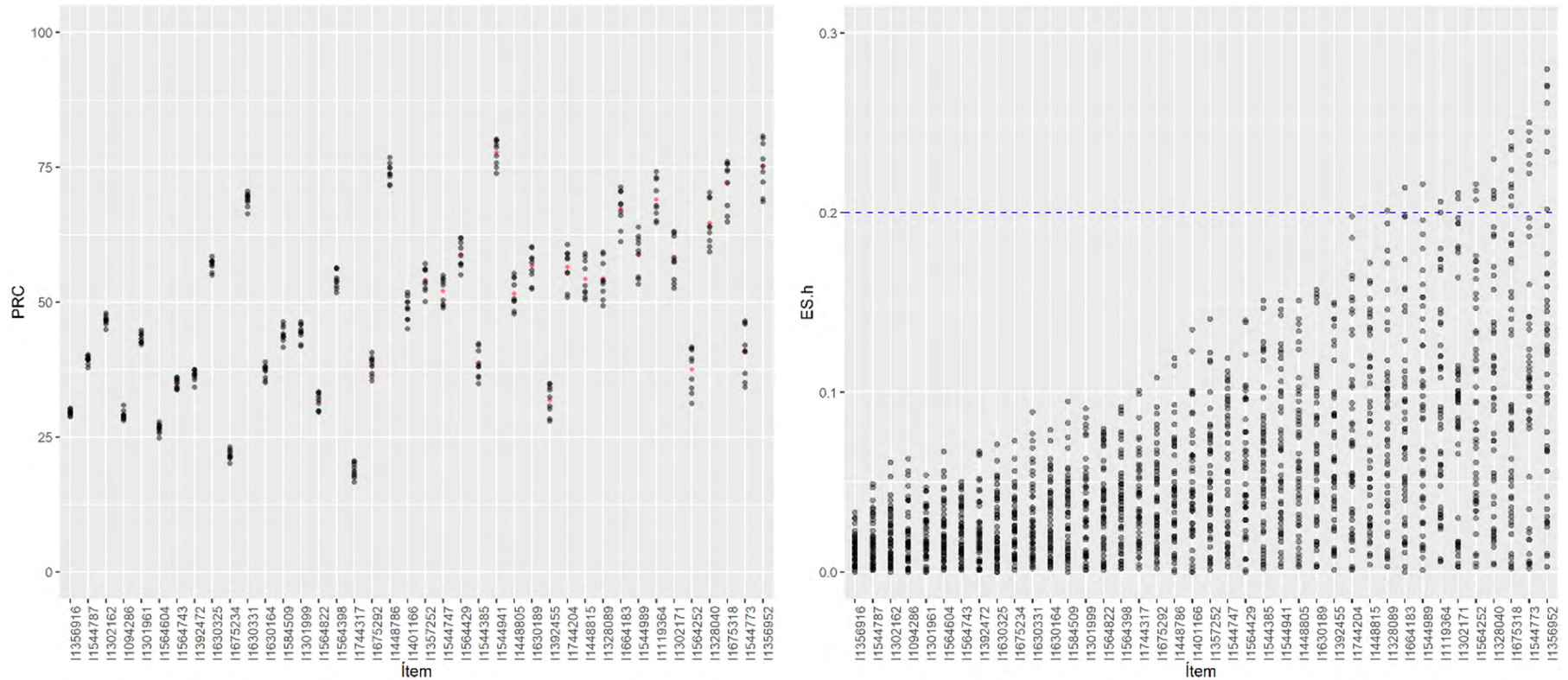
Figura 3. Resultados DIF Saber TyT 2020-1



3.1.3. Comparación de PRC de ítems

Para comprender en mayor detalle los resultados que se obtienen de la comparación de PRC de ítems, se presenta la **Figura 4** con los resultados para la prueba de Razonamiento Cuantitativo. En el panel izquierdo se presenta para cada ítem los valores del PRC en las distintas formas y sesiones de aplicación, ordenándolos del que tiene menor variabilidad a mayor variabilidad. En el panel derecho, los puntos corresponden al valor absoluto del TE al hacer todas las comparaciones dos a dos del PRC entre formas y sesiones para cada ítem. En esta, un ítem con diferencias grandes de PRC entre sesiones o formas está relacionado con valores más grandes del TE. En el panel derecho de la figura se observa que la mayoría de los ítems tienen un TE despreciable en la totalidad de sus comparaciones del PRC, y que solo 10 ítems, los cuales se ubican hacia la derecha, presentan alguna comparación clasificada con TE pequeño. De estos, sólo 3 ítems presentan 5 o más comparaciones en esta clasificación.

Figura 4. Resultados al comparar el PRC entre formas y sesiones para Razonamiento cuantitativo Saber TyT 2020-1.



A partir de lo anterior, en la **Tabla 5** se presenta un resumen de los resultados para la prueba Saber TyT en su aplicación 2020-1, en esta se observa qué un porcentaje alto (mayor al 90%) de las comparaciones se clasifican en tamaño del efecto despreciable para las pruebas. Respecto al número de ítems con alguna comparación del TE clasificada en pequeñas, se observa un total de 13 ítems, para los cuales 5 presentan 5 o más TE clasificados en pequeños. Para los demás ítems se obtuvieron diferencias despreciables, lo cual indica que su comportamiento fue homogéneo en todas las sesiones y formas aplicadas.

3.1.4. Calificación de la prueba electrónica

En la **Tabla 6** se presenta la comparación de los puntajes promedio y desviaciones estándar entre las aplicaciones de 2018-1, 2019-1 y 2020-1 y se observa que, en las pruebas de Inglés y Lectura crítica, el promedio del puntaje tiende a ser similar entre las aplicaciones del 2018-1 y 2019-1; mientras que entre estos dos periodos el promedio disminuye 4,9 puntos en la prueba de Competencias ciudadanas y 2,6 puntos en la de Razonamiento cuantitativo. Entre las aplicaciones de 2019-1 y 20202-1 el promedio del puntaje tuvo cambios leves en todas las pruebas, lo cual es un buen indicio de comparabilidad entre formatos.

Respecto a la desviación estándar, esta aumentó en mayor medida en las pruebas de Inglés y Lectura crítica entre la aplicación de 2019-1 y 2020-1. Este mismo comportamiento se evidencia en las pruebas de Competencias ciudadanas y Razonamiento cuantitativo, sin embargo, este aumento fue leve.

Tabla 5. Comparación de estudiantes que presentaron Saber TyT en 2018-1 (prueba en papel) y en 2020-1 (prueba en computador), la diferencia entre las dos aplicaciones y el tamaño del efecto (TE)

Prueba	No. De comparaciones (Estadísticas TE)					No. de ítems		
	Análisis	TE		% TE Despreciable	% TE Pequeño	Análisis	Con 1 o más TE Pequeño	Con 5 o más TE Pequeño
		Despreciable	Pequeño					
Lectura Crítica	1800	1787	13	99	1	40	1	1
Razonamiento Cuantitativo	1800	1766	34	98	2	40	10	3
Competencias Ciudadanas	1800	1790	10	99	1	40	2	1
Inglés	370	370	0	100	0	65	0	0
Total	5770	5713	57	99	1	185	13	5

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

PRUEBA	Promedio puntaje			Desviación		
	2018-1	2019-1	2020-1	2018-1	2019-1	2020-1
Competencias ciudadanas	99,7	94,8	95,1	23,0	23,8	24,5
Inglés	100,2	100,0	98,9	21,4	22,1	28,0
Lectura crítica	99,7	99,1	98,6	21,6	21,1	26,5
Razonamiento cuantitativo	94,0	91,4	89,6	21,9	21,4	22,4

Adicionalmente, la **Figura 5** permite visualizar la información respecto a los puntajes para los tres periodos. Teniendo en cuenta que la prueba se aplicó en papel en 2018 y 2019, mientras que en 2020 fue electrónica, tanto la **Figura 5** como la **Tabla 6** permiten comparar el comportamiento de las pruebas en diferentes formatos. En estos resultados se esperaría que las distribuciones no se movieran entre aplicaciones o que sus cambios fueran pequeños. Al revisar esto con la **Figura 5**, se identifica que las distribuciones por aplicación son similares, excepto en Competencias ciudadanas, donde la distribución de la aplicación 2020-1 se concentra levemente hacia la derecha. Adicionalmente, en inglés se presentó un alto nivel de omisiones, lo cual se evidencia en la agrupación de datos que se presenta en la parte inferior de la distribución.

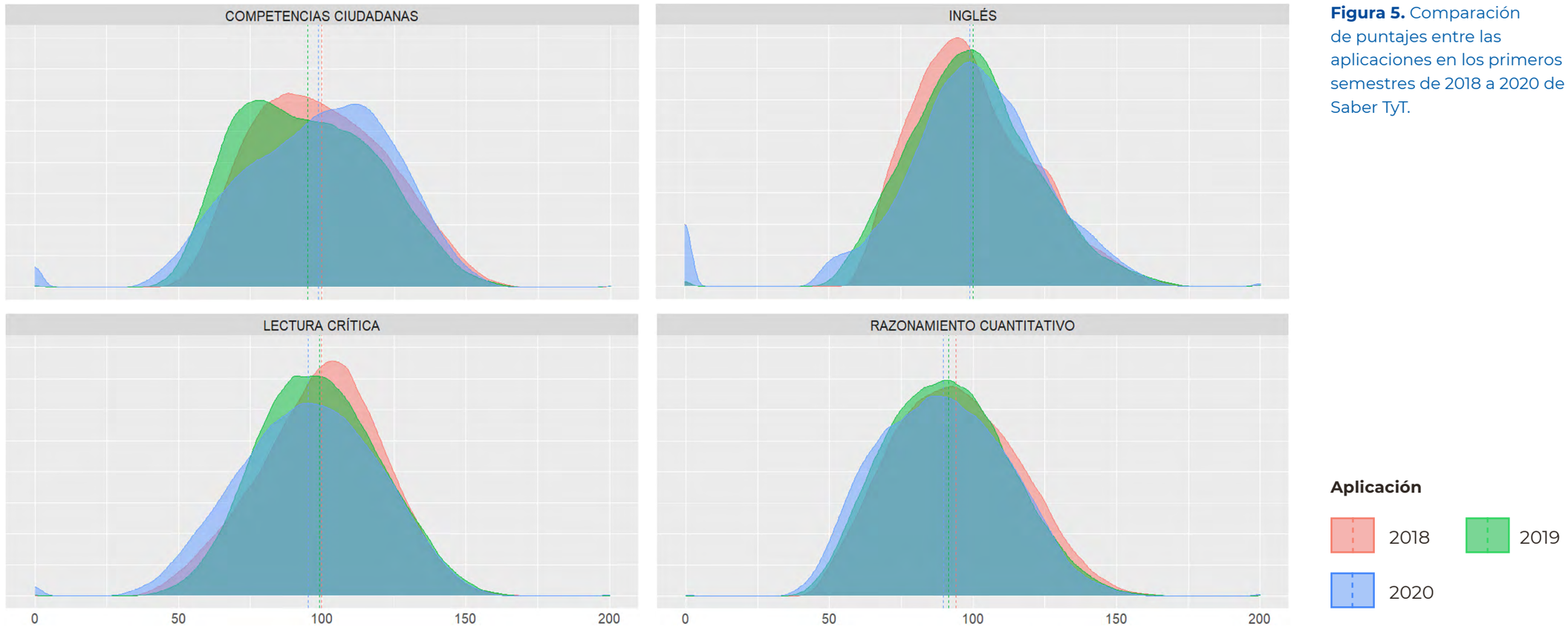


Figura 5. Comparación de puntajes entre las aplicaciones en los primeros semestres de 2018 a 2020 de Saber TyT.

3.2. Análisis de resultados pruebas Saber TyT 2020-3

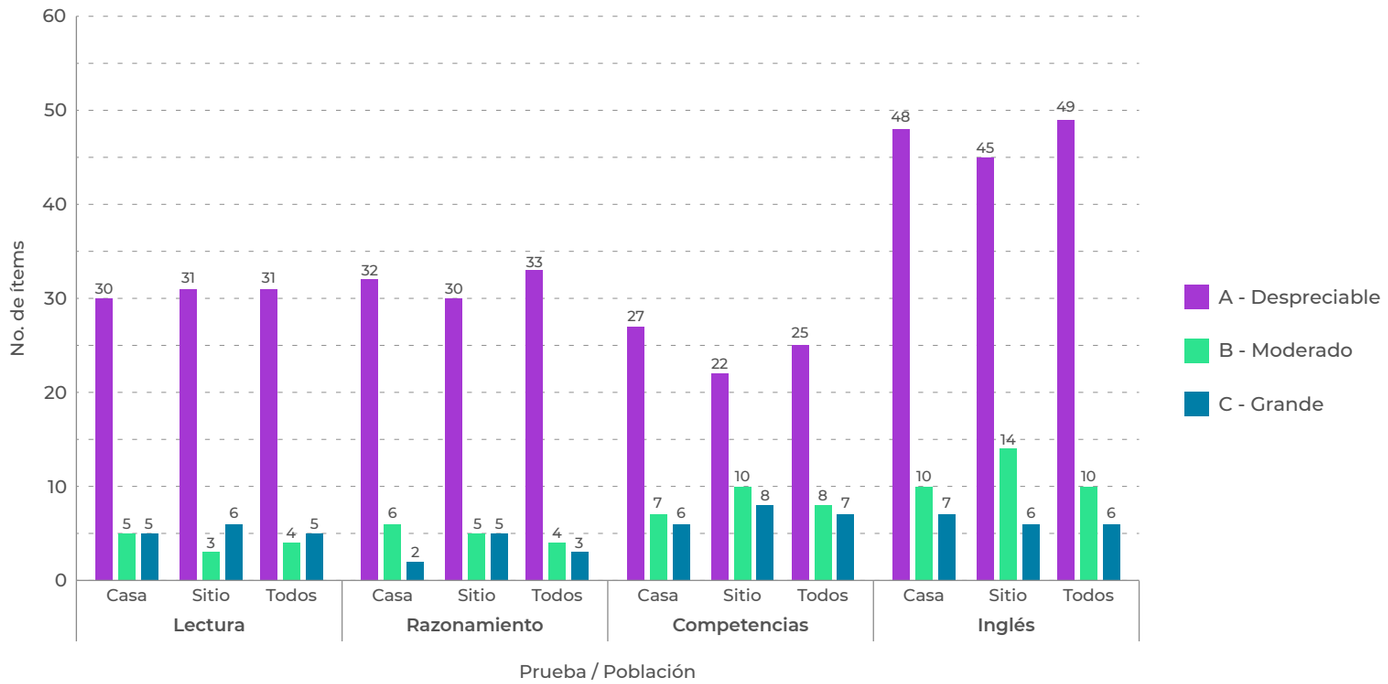
En la aplicación de Saber TyT 2020-3 se ofrecieron las dos modalidades de presentación para los evaluados, es decir, en casa y en sitio. Por lo tanto, en esta ocasión, no estuvo excluida la población de estudiantes que tenían dificultades para acceder a un computador y conexión a internet en casa, como fue el caso de Saber TyT 2020-1. Dado que el grupo de estudiantes que presentaron la versión electrónica de Saber TyT 2020-3 incluyó a toda la población gracias a la aplicación en sitio, se pudo considerar que el grupo de estudiantes es equivalente a la población típica que presentaba la prueba en papel antes de la pandemia, de manera que se procedió directamente al análisis de DIF sin pasar por la comparación de poblaciones, como sí fue necesario para Saber TyT 2020-1.

3.2.1. Análisis de DIF

Para analizar si existían comportamientos diferenciales en los ítems a partir del formato de aplicación, se realizaron tres tipos de análisis: **1)** análisis de DIF para los ítems calibrados de los evaluados en casa contra los resultados históricos (en papel), **2)** análisis de DIF para los ítems calibrados de los evaluados en sitio contra los resultados históricos y **3)** análisis de DIF para los ítems calibrados con toda la población (sitio y casa) contra los resultados históricos. En los tres análisis las calibraciones históricas se consideran las del grupo focal.

Teniendo en cuenta lo anterior, en la Figura 6 se presenta el número total de ítems por modalidad de aplicación y prueba, en las tres categorías de tamaño del efecto de DIF resultante de clasificar el índice NcDIF. Se observa que hay un total de 40 ítems para las pruebas de lectura, razonamiento y competencias ciudadanas, mientras que para inglés el total es de 65 ítems. De acuerdo con estos valores absolutos, y referente a las comparaciones de resultados entre casa y sitio, se observa en la Figura que la prueba con mayor número de ítems de forma que presentan posible DIF C para la población evaluada en sitio es la de Competencias ciudadanas, mientras que Lectura crítica tuvo menor cantidad de ítems con posible DIF.

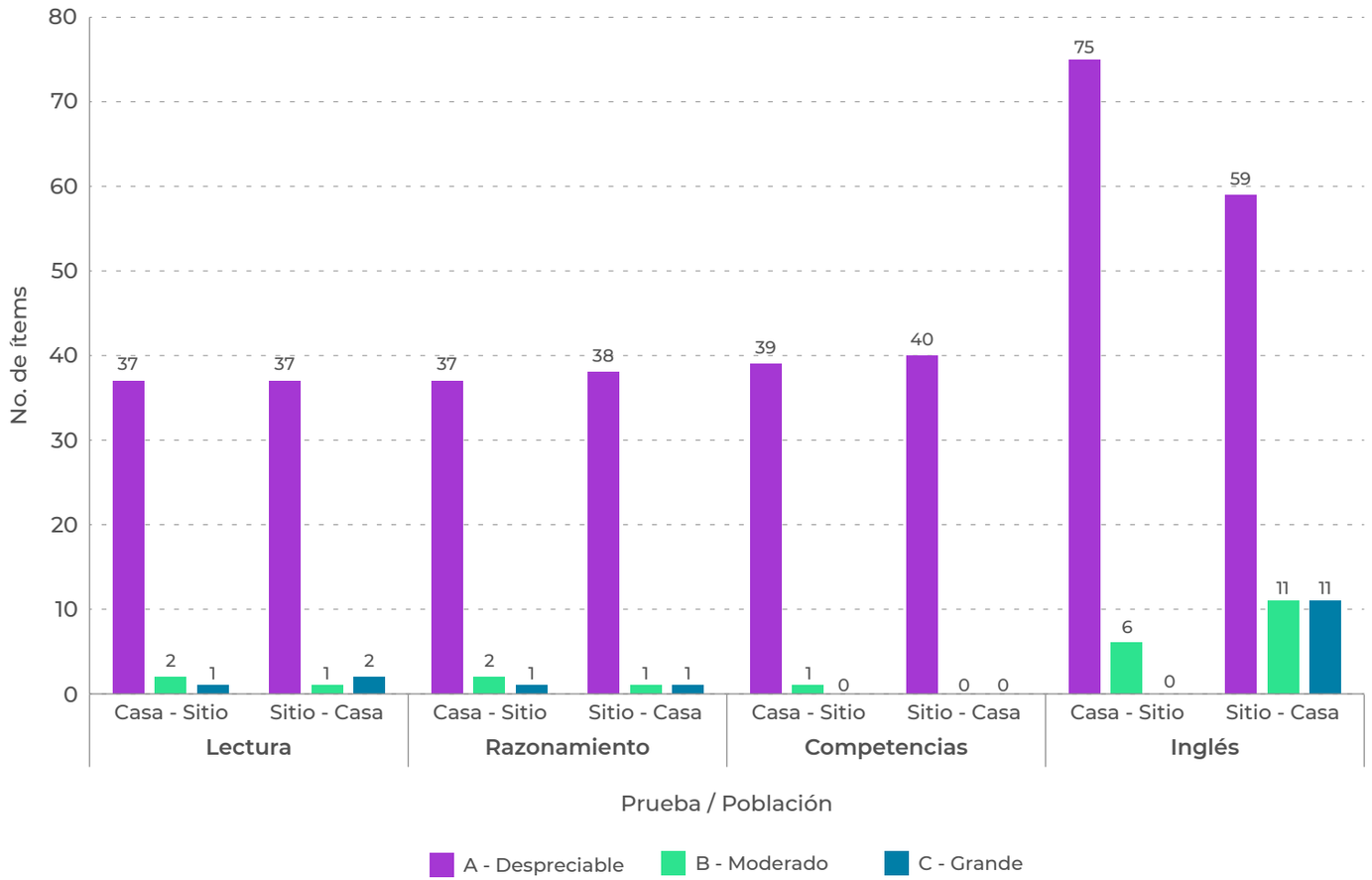
Figura 6. Comparación de los resultados del análisis de DIF por los análisis de aplicación en casa, sitio y el total de la población que presentó el examen Saber TyT 2020-3.



También era de interés verificar si existió algún tipo de DIF entre los resultados obtenidos con la población evaluada en casa en contraste con los de la población evaluada en sitio. La **Figura 7** representa los conteos de ítems clasificados en las categorías de DIF, en esta el análisis "Casa-Sitio" define como grupo focal los evaluados en casa, mientras que en el análisis "Sitio-Casa" se define el grupo focal los evaluados en sitio.

En la **Figura 7** se observan resultados consistentes para las pruebas de Lectura Crítica, Razonamiento Cuantitativo y Competencias Ciudadanas, con una mínima cantidad de ítems en las categorías de DIF B o C al comparar la aplicación en casa con la aplicación en sitio. Para el caso de la prueba de inglés, el panorama cambia, y la cantidad de ítems aumenta considerablemente, en particular, cuando el grupo focal cambia de casa a sitio, se observa un aumento de 6 a 22 ítems. Teniendo en cuenta lo anterior, para realizar el seguimiento a los ítems en categorías B y C se utiliza la identificación de ambos análisis (Casa-Sitio y Sitio-Casa), en el que se observa que varios ítems se identifican en ambos.

Figura 7. Comparación de resultados de DIF entre las aplicaciones de casa y sitio para el examen Saber TyT 2020-3..



3.2.2. Análisis de DIF complementario

A continuación, como se explicó en la sección de metodología, se presentan los resultados de los análisis de DIF complementarios para el examen Saber TyT 2020-3, en el cual se muestra la clasificación del tamaño del efecto calculado a partir de los estadísticos Mantel-Haenszel (MH) y de regresión logística (Log).

En la **Tabla 7** se observa que ambos estadísticos marcan pocos ítems con TE de DIF no despreciable, cuatro en total; y en particular el estadístico Log marca un único ítem, perteneciente a la prueba de Razonamiento Cuantitativo, el cual también es marcado por el estadístico MH. Esos resultados, salvo para la prueba de Inglés, son coherentes con los análisis realizados a partir del NcDIF, y nos permiten afirmar que no se evidencian diferencias grandes entre las respuestas a los ítems entre los dos tipos de aplicación para las pruebas de Lectura Crítica, Competencias Ciudadanas y Razonamiento Cuantitativo.

Tabla 7. Clasificación por tamaño del efecto de DIF según metodología Saber TyT 2020-3.

Prueba	# ítems	Clasificación MH			Clasificación Log			Clasificación Máxima			Marcados con DIF
		# ítems A	# ítems B	# ítems C	# ítems A	# ítems B	# ítems C	# ítems A	# ítems B	# ítems C	
Competencias ciudadanas	40	40	0	0	40	0	0	40	0	0	0
Inglés	90	89	1	0	90	0	0	89	1	0	1
Lectura crítica	40	39	1	0	40	0	0	39	1	0	1
Razonamiento cuantitativo	40	38	1	1	39	1	0	38	1	1	2

3.2.3. Comparación de PRC de ítems

En la **Tabla 8** se presentan los resultados de la comparación del PRC de los ítems en la aplicación Saber TyT 2020-3, en esta se observa qué un porcentaje alto (mayor al 90%) de las comparaciones se clasifican en tamaño del efecto despreciable para las pruebas salvo en competencias en el que este porcentaje es solo del 85%. Respecto al número de ítems con alguna comparación TE clasificada en pequeñas se observa un total de 75 ítems, para los cuales 39 presentan 5 o más TE clasificadas en pequeñas. La mayoría de los ítems son de competencias (12) e inglés (22) y se les hizo seguimiento en los análisis de ítem y de DIF. Para el resto de los ítems se obtuvieron diferencias despreciables, lo cual indica que su comportamiento fue homogéneo en las sesiones y tipo de aplicación.

3.2.4. Calificación de la prueba electrónica

Para realizar los resultados presentados en este apartado, se tomaron las aplicaciones denominadas por el instituto como 2018-3, 2019-5 y 2020-3, las cuales corresponden al segundo periodo de los años mencionados. De acuerdo con la **Tabla 9**, el promedio del puntaje y la desviación estándar ha variado entre años. Entre 2018 y 2019, el promedio de puntaje disminuyó 8,5 puntos en Competencias ciudadanas, mientras que el promedio de las demás pruebas varió levemente. Y entre el 2019 y 2020, el promedio del puntaje de todas las pruebas subió entre 1 a 3 puntos, aproximadamente. Adicionalmente, en los periodos comparados la desviación estándar de todas las pruebas ha tenido variaciones mínimas.

Tabla 8. Resultados prueba Saber TyT 2020-3.

Prueba	No. De comparaciones (Estadísticas TE)					No. de ítems		
	Análisis	TE		% TE Despreciable	% TE Pequeño	Análisis	Con 1 o más TE Pequeño	Con 5 o más TE Pequeño
		Despreciable	Pequeño					
Lectura Crítica	840	821	19	98	2	40	7	1
Razonamiento Cuantitativo	840	804	36	96	4	40	15	1
Competencias Ciudadanas	840	716	124	85	15	40	24	15
Inglés	3.240	2.954	286	91	9	90	29	22
Total	5.760	5.295	465	92	8	210	75	39

Tabla 9. Promedio del puntaje y desviación estándar para los módulos generales, según el periodo de aplicación.

PRUEBA	Promedio puntaje			Desviación		
	2018	2019	2020	2018	2019	2020
Competencias ciudadanas	99,3	90,8	94,0	23,7	25,4	23,0
Inglés	100,1	100,6	101,5	22,3	21,9	23,3
Lectura crítica	100,2	96,8	98,6	21,5	22,7	25,7
Razonamiento cuantitativo	91,2	88,1	89,6	22,8	20,9	21,2

Al graficar los puntajes de los evaluados por cada aplicación, se identifica que el comportamiento de las distribuciones de cada una, son similares entre sí. Al igual que en el resumen presentado para la aplicación TyT 2020-1, la distribución de la aplicación 2020-2 de la prueba de Competencias ciudadanas se concentra levemente hacia la derecha. Por otra parte, en la prueba de Inglés se disminuye el inconveniente de las omisiones que se presentó en la primera aplicación (Ver **Figura 8**).

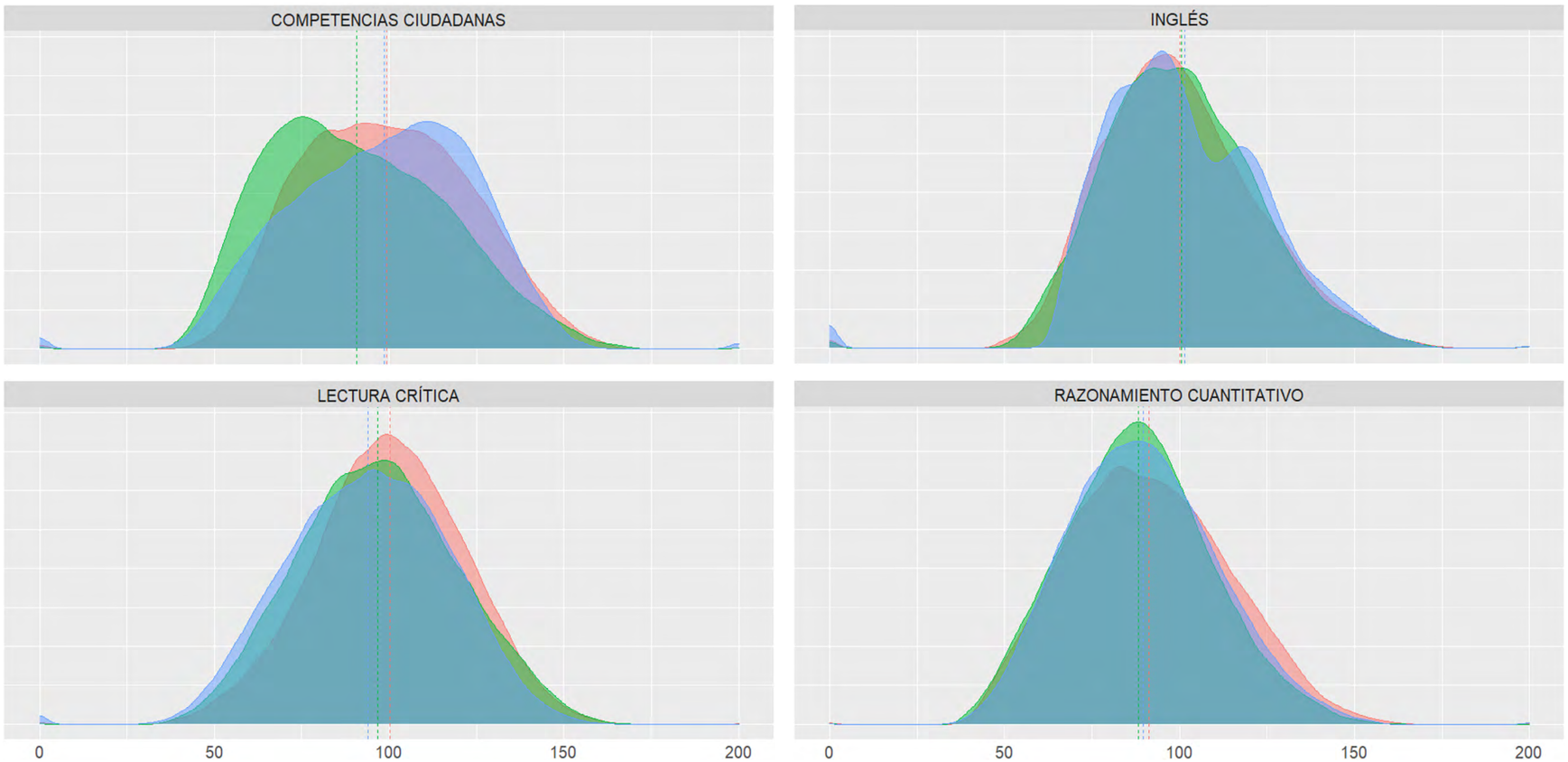


Figura 8. Comparación de puntajes entre las aplicaciones en los segundos semestres 2018 a 2020 de Saber TyT.

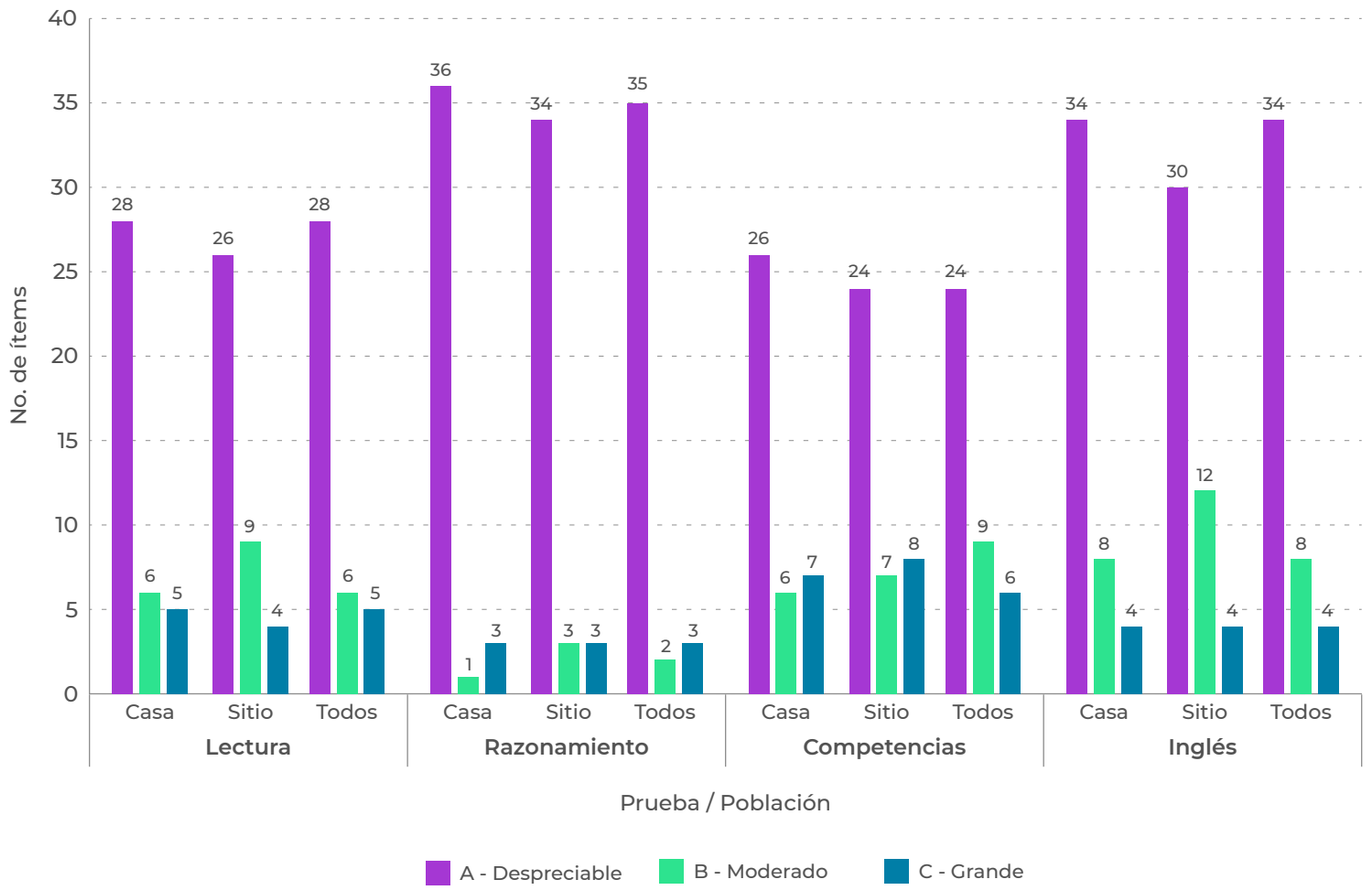
3.3. Análisis de resultados pruebas Saber Pro 2020-3

De manera similar a Saber TyT 2020-3, en la aplicación de Saber Pro-2020-3 se ofrecieron las dos modalidades de presentación para los evaluados, es decir, en casa y en sitio. Por lo tanto, no se vio restringida la población de estudiantes que tenían dificultades para acceder a computador y conexión a internet en casa, y se asumió que el grupo de evaluados era equivalente a la población típica que presentaba la prueba en papel antes de la pandemia. Tal como ocurrió con Saber TyT 2020-3, no se realizó una comparación de poblaciones, sino que se procedió directamente al análisis de DIF.

3.3.1. Análisis de DIF

La **Figura 9** presenta el número de ítems obtenidos en cada una de las categorías de clasificación de DIF para los tres tipos de análisis realizados en cada una de las pruebas evaluadas. Se observa que hay un total de 39 ítems para la prueba de lectura y competencias ciudadanas, 40 ítems para las pruebas de razonamiento, mientras que para inglés el total es de 46 ítems. En general, la mayoría de los ítems se encuentran clasificados en la categoría A de DIF despreciable para todas las pruebas evaluadas. En particular, las pruebas de Razonamiento Cuantitativo y la de Inglés son las que tienen mayor cantidad de ítems en esta categoría.

Figura 9. Comparación de resultados de DIF entre las aplicaciones de casa y sitio para el examen Saber Pro.

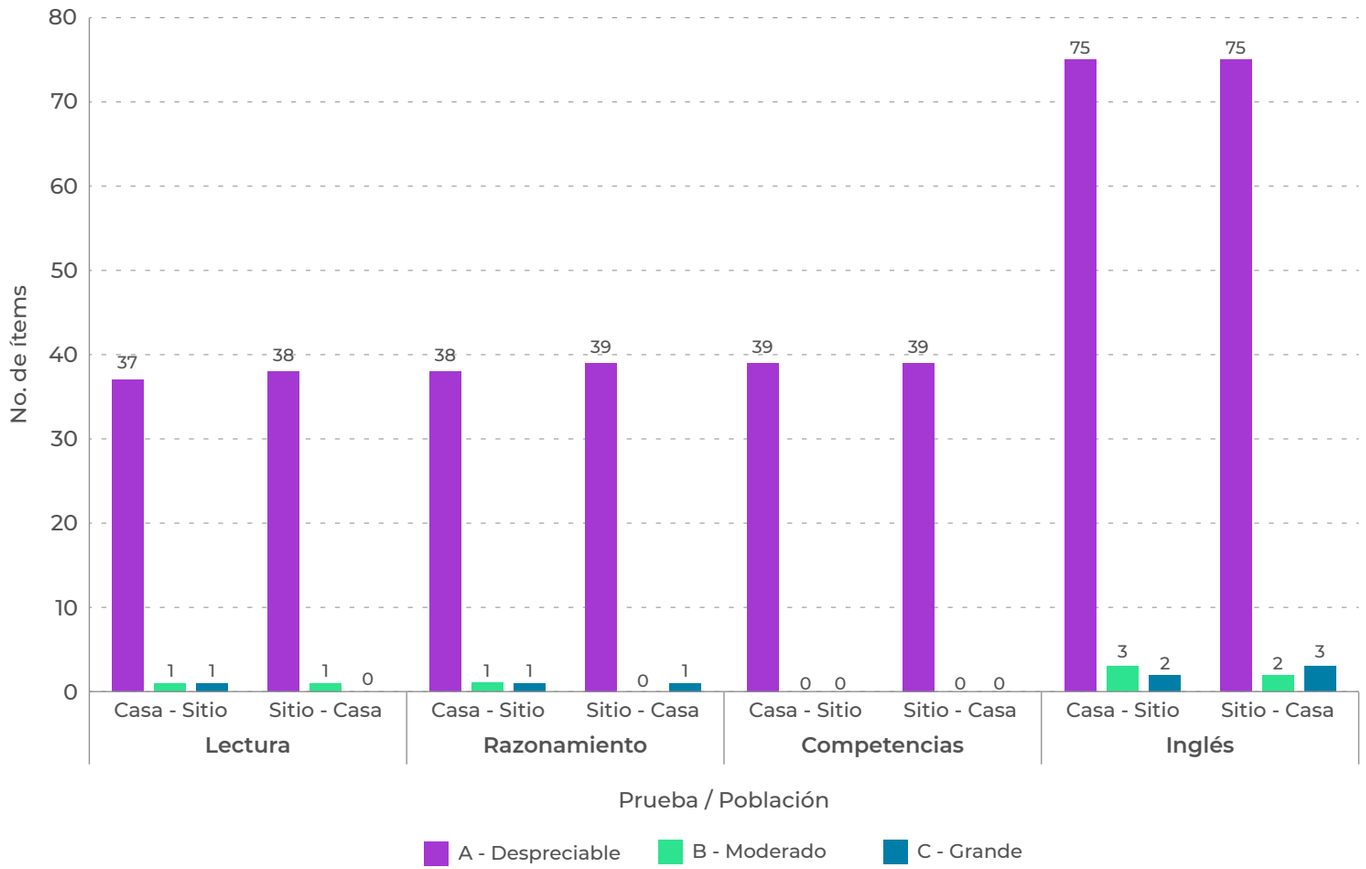


Frente a las comparaciones entre casa y sitio, se observa que para todas las pruebas evaluadas hay mayor cantidad de ítems categorizados en B o C en la aplicación en sitio, observándose la mayor diferencia en la prueba de inglés con un valor de 4 ítems más en estas categorías. Mientras que para las demás pruebas evaluadas las diferencias corresponden a 2 ítems[1].

En cuanto al análisis puntual de comparación de parámetros de ítems obtenidos entre la población evaluada en casa y la evaluada en sitio, la **Figura 10** representa los conteos de ítems clasificados en las categorías de DIF. La información se presenta teniendo en cuenta el cambio en el grupo focal o referencia.

En general, se observan resultados consistentes para todas las pruebas evaluadas, con una mínima cantidad de ítems en las categorías de DIF B o C. Nuevamente, los resultados de la prueba de inglés son un poco distintos respecto a las demás pruebas, pero para este caso, solo se hallaron 5 ítems con DIF B o C. Por lo tanto, hay un buen número de ítems sin DIF con los que se puede equiparar la aplicación electrónica de 2020-3 con las calibraciones del histórico para tener los puntajes en la misma escala de la línea base.

Figura 10. Comparación de resultados de DIF entre las aplicaciones de casa y sitio para el examen Saber Pro.



1 Si desea ver los anteriores análisis desagregados por tipo de DIF (Uniforme y No uniforme) el lector puede remitirse a la sección de Anexos.

3.3.2. Análisis de DIF complementario

A continuación, como se explicó en la sección de metodología, se presentan los resultados de los análisis de DIF complementario para el examen Saber Pro 2020-3, en el cual se muestra la clasificación del tamaño del efecto calculado a partir de los estadísticos Mantel-Haenszel (MH) y de regresión logística (Log).

En analogía con lo observado para el examen Saber TyT, nuevamente se observa que ambos estadísticos marcan pocos ítems con TE de DIF no despreciable, cuatro en total, de los cuales tres pertenecen a la prueba de Inglés y el restante a la de Razonamiento Cuantitativo. Adicionalmente, el estadístico Log detectó un único ítem en la prueba de Inglés, el cual también fue detectado por el estadístico MH. Nuevamente, estos resultados son coherentes con los análisis realizados a partir del NcDIF, y nos permiten afirmar que no se evidencian diferencias grandes entre las respuestas a los ítems entre los dos tipos de aplicación para las pruebas del examen Saber Pro 2020-3 (Ver **Tabla 10**)

Tabla 10. Clasificación por tamaño del efecto de DIF según metodología Saber Pro 2020-3.

Prueba	# ítems	Clasificación MH			Clasificación Log			Clasificación Máxima			Marcados con DIF
		# ítems A	# ítems B	# ítems C	# ítems A	# ítems B	# ítems C	# ítems A	# ítems B	# ítems C	
Competencias	40	40	0	0	40	0	0	40	0	0	0
Inglés	90	89	1	2	89	1	0	87	1	2	3
Lectura	40	40	0	0	40	0	0	40	0	0	0
Razonamiento	40	39	1	0	40	0	0	39	1	0	1

3.3.3. Comparación de PRC de ítems

En la **Tabla 11** se encuentran los resultados de la comparación de los PRC de los ítems para la prueba Saber Pro 2020-3, en esta se observa que un porcentaje alto (mayor al 90%) de las comparaciones se clasifican en tamaño del efecto despreciable para las pruebas salvo en Inglés, en el que este porcentaje es solo del 82%. Respecto al número de ítems con alguna comparación TE clasificada en pequeñas se observa un total de 91 ítems, de los cuales 43 presentan 5 o más TE clasificadas en pequeñas. La mayoría de los ítems son de Inglés (36) y se les hizo seguimiento en los análisis de ítem y de DIF. Para los demás ítems, se obtuvieron diferencias despreciables, lo cual indica que su comportamiento fue homogéneo en las sesiones y tipo de aplicación.

3.3.4. Calificación de la prueba electrónica

Para las comparaciones de los puntajes en las aplicaciones de Saber Pro se tomaron las aplicaciones de como 2020, 2019 y 2018. De acuerdo con la **Tabla 12**, entre el 2018 y 2019 el promedio del puntaje presentó una variación leve en todas las pruebas; mientras que entre el 2019 y 2020 el promedio del puntaje aumentó en todas las pruebas, sin embargo, este cambio es mayor en la prueba de competencias ciudadanas ya que incrementó 10 puntos. Respecto a la desviación estándar, para todas las pruebas se presentaron variaciones mínimas en los periodos comparados.

Tabla 11. Resultados de la comparación PRC Saber Pro 2020-3.

Prueba	No. De comparaciones (Estadísticas TE)					No. de ítems		
	Análisis	TE		% TE Despreciable	% TE Pequeño	Análisis	Con 1 o más TE Pequeño	Con 5 o más TE Pequeño
		Despreciable	Pequeño					
Lectura Crítica	840	834	6	99	1	40	3	0
Razonamiento Cuantitativo	840	784	56	93	7	40	20	4
Competencias Ciudadanas	840	783	57	93	7	40	21	3
Inglés	3219	2638	581	82	18	90	47	36
Total	5.760	5.295	465	92	8	210	75	39

Tabla 12. Promedio del puntaje y desviación estándar para los módulos generales, según el periodo de aplicación.

PRUEBA	Promedio puntaje			Desviación		
	2018	2019	2020	2018	2019	2020
Inglés	150,7	151,7	156,0	32,2	32,2	34,1
Lectura crítica	150,0	148,9	151,6	30,9	31,4	30,2
Competencias ciudadanas	142,1	142,4	152,4	33,1	33,3	32,1
Razonamiento cuantitativo	149,6	146,4	148,6	31,0	31,7	32,2

Adicionalmente, al igual que las aplicaciones de Saber TyT, en la **Figura 11** se identifica que la distribución de la prueba de Competencias ciudadanas también se concentra levemente hacia la derecha en la aplicación 2020-3. Sin embargo, la prueba de Inglés no presenta el problema de omisiones mencionado en las pruebas Saber TyT.

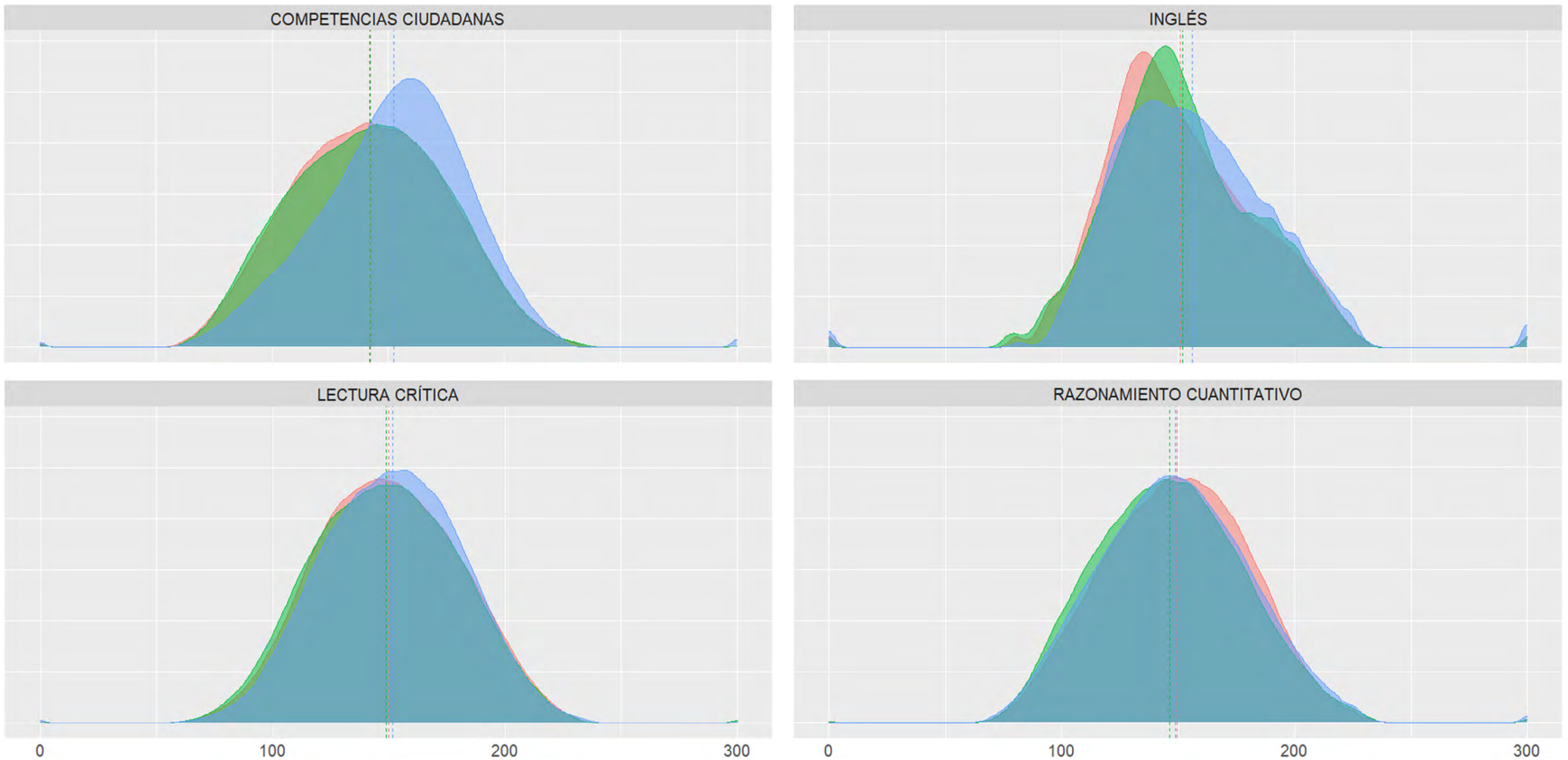


Figura 11. Comparación de puntajes entre 2018 y 2020 de Saber TyT 2020-1..

Examen Saber Pro

Examen Saber TyT

04

Conclusiones

Dada la situación generada por la pandemia en el 2020, no fue posible aplicar las pruebas Saber TyT y Pro de la manera convencional en ese momento. Por lo tanto, aprovechando las experiencias previas que tenía el Icfes con pruebas electrónicas, se decidió implementar este formato para que los estudiantes de educación superior pudieran presentar la prueba y cumplir con su requisito de grado. Además, esto constituyó una oportunidad para que el Instituto pudiera hacer una observación a gran escala con estudiantes de educación superior para analizar la comparabilidad de la prueba en formato electrónico con las pruebas en papel.

Como se presentó en secciones anteriores, se siguieron varios pasos para el análisis de comparabilidad de las pruebas en formato electrónico en 2020 con los resultados de las pruebas en papel anteriormente aplicadas. En primer lugar, dado que la aplicación para Saber TyT 2020-1 se llevó a cabo solo en casa con estudiantes que tuvieran computador y conexión a internet, se analizó si esta población era comparable con toda la población que presentaba las pruebas en papel anteriormente. Dado que se observó comparabilidad entre poblaciones, se pudo proceder con los siguientes análisis. Para Saber TyT 2020-3 y Saber Pro 2020-3, se ofreció aplicar la prueba en sitio también, especialmente para los estudiantes sin las condiciones para aplicar la prueba electrónica desde casa, de manera que al evaluar toda la población de estudiantes, no fue necesario analizar la comparabilidad con la población que presentaba las pruebas anteriormente en papel.

Al tener poblaciones comparables, se procedió a realizar análisis de DIF entre la versión electrónica de la prueba y las calibraciones históricas que se tenían de la prueba en papel. En general, se observó que el número de ítems con DIF era pequeño, de manera que se tenía un buen número de ítems sin DIF para equiparar la prueba electrónica con la escala histórica que se tenía en papel desde la línea base. Esto permitió mantener la misma escala de calificación para las pruebas electrónicas. Estos resultados confirman las conclusiones de Choi & Tinkler (2002), en cuanto a que estudiantes en grados de escolaridad más altos, presentan mayor comparabilidad al presentar la prueba en papel y en computador.

Por otra parte, en los análisis de comparación de PRC entre formas, sesiones y modalidad de aplicación (casa, sitio), la mayoría de los ítems presentaron TE despreciables y se identificaron unos pocos con TE pequeños, las cuales no son diferencias de gran magnitud que indiquen comportamientos irregulares. Por lo tanto, en general se observaron comportamientos homogéneos entre formas, sesiones y modalidad de aplicación.

Respecto a la comparación de los promedios de puntajes por aplicaciones, se observa que no hay un cambio grande en el comportamiento de la distribución, por lo que esto puede evidenciar que el cambio de formato de aplicación en las pruebas no genera grandes diferencias en los resultados de los evaluados.

Por último, también se identifica que la variabilidad de los promedios es leve entre las aplicaciones tanto de las pruebas Saber TyT como Saber Pro. Todo esto indica que se puede considerar que el cambio de formato pasando a pruebas electrónicas permite mantener la misma escala de calificación que la que se tenía cuando se aplicaban las pruebas en papel.

05.

Referencias

Referencias

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP. 39.

Choi, S. W., & Tinkler, T. (2002). Evaluating comparability of paper-and-pencil and computer-based assessment in a K–12 setting. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), New Orleans, LA. (s. f.). https://nceo.info/references/paper-conference/10706

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155–159. doi:10.1037/0033- 2909.112.1.155

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

Congreso de la República de Colombia. (14 de octubre 2009). Decreto 3963 de 2009. DO: 4650.

Congreso de la República de Colombia. (12 de noviembre de 2020). Resolución 530 de 2020. [Derogada]. DO: 51.497.

Diamond, A., & Sekhon, J. S. (2013). Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. Review of Economics and Statistics, 95(3), 932-945.

Hardcastle, J., Herrmann-Abell, C. F., & DeBoer, G. E. (2017). Comparing Student Performance on Paper-and-Pencil and Computer-Based-Tests. Grantee Submission.

Instituto Colombiano para la Evaluación de la Educación (2019a). Informe de Gestión.Vigencia 2019, Icfes. Bogotá, Colombia: autor. Recuperado de https://www.icfes.gov.co/documents/39286/2327961/Informe+de+gestion+2019.pdf/14be4033-d55f-fadd-c396-fa80e9ea939b?version=1.0&t=1647984308521

Instituto Colombiano para la Evaluación de la Educación (2019b). ¿En qué consiste la aplicación de Pre SABER 11° en versión adaptativa (CAT)?. Saber al detalle, 3 (6), [1-8]. Recuperado de https://www.icfes.gov.co/documents/39286/2231027/Edicion+6+-+boletin+saber+al+detalle+.pdf/1db5914f-5496-9cd1-2c69-7efb79c27faa? version=1.0&t=1647958805503

Instituto Colombiano para la Evaluación de la Educación (2020a). Informe nacional de resultados del examen Saber Pro, Icfes. Bogotá, Colombia: autor.

Instituto Colombiano para la Evaluación de la Educación (2020b). Informe nacional de resultados del examen Saber TyT, Icfes. Bogotá, Colombia: autor.

Instituto Colombiano para la Evaluación de la Educación (2020c). Informe de Gestión.Vigencia 2020, Icfes. Bogotá, Colombia: autor. Recuperado de https://www.icfes.gov.co/documents/39286/2327961/Informe+de+gestion+2020.pdf/a1d68b87-d5d3-bbfc-dd81-e945fbda82de?version=1.0&t=1647984311183

Instituto Colombiano para la Evaluación de la Educación (2022). Estudio cuasiexperimental para establecer las diferencias entre métodos de aplicación de pruebas estandarizadas en la República Dominicana, Icfes. Bogotá, Colombia: autor.

Jodoin, M. G. and Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. Applied Measurement in Education, 14, 329-349. doi: 10.1207/S15324818AME1404_2

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

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. Biometrika, 78, 691-692. doi: 10.1093/biomet/78.3.691

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.

Quintero, A., Shavelson, R., Rodríguez, A.,Duplat, R., y Calderón, A. (2022). On the comparability of scores from paper- and computer-based achievement tests: Challenges and findings from quasi-experiments. Artículo en preparación.

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

.90

Anexos

Examen Saber Pro

Examen Saber TyT

Resultados Análisis de DIF
Saber TyT 2020-1 desagregado por tipo de DIF

Prueba	NCDIF	No DIF	No Uniforme	Uniforme	Total
Inglés	A	2	26	9	37
	B	2	4	2	8
	C	0	3	5	8
Competencias Ciudadanas	A	9	11	4	24
	B	1	6	4	11
	C	1	4	0	5
Lectura Crítica	A	5	21	4	30
	B	2	3	2	7
	C	1	1	1	3
Razonamiento Cuantitativo	A	6	20	3	29
	B	0	4	4	8
	C	0	0	3	3

Resultados Análisis de DIF
Saber TyT 2020-3 desagregado por tipo de DIF

Prueba	Tipo	NCDIF	No DIF	No Uniforme	Uniforme	Total
Inglés	Casa	A	10	29	9	48
		B	1	6	3	10
		C	0	4	3	7
	Sitio	A	8	30	7	45
		B	1	8	5	14
		C	0	5	1	6
	Todos	A	8	30	11	49
		B	0	5	5	10
		C	1	4	1	6
Competencias Ciudadanas	Casa	A	6	16	5	27
		B	1	4	2	7
		C	0	4	2	6
	Sitio	A	6	14	2	22
		B	1	6	3	10
		C	0	6	2	8
	Todos	A	5	16	4	25
		B	1	5	2	8
		C	0	5	2	7

Prueba	Tipo	NCDIF	No DIF	No Uniforme	Uniforme	Total
Lectura Crítica	Casa	A	8	20	2	30
		B	0	2	3	5
		C	3	2	0	5
	Sitio	A	11	15	5	31
		B	1	0	2	3
		C	3	3	0	6
	Todos	A	9	15	7	31
		B	0	2	2	4
		C	3	2	0	5
Razonamiento Cuantitativo	Casa	A	10	15	7	32
		B	0	2	4	6
		C	0	1	1	2
	Sitio	A	9	12	9	30
		B	1	2	2	5
		C	1	2	2	5
	Todos	A	11	13	9	33
		B	0	2	2	4
		C	0	2	1	3

Resultados Análisis de DIF
Saber Pro 2020-3 desagregado por tipo de DIF

Prueba	Tipo	NCDIF	No Uniforme	Uniforme	Total
Inglés	Casa	A	22	12	34
		B	4	4	8
		C	1	3	4
	Sitio	A	20	10	30
		B	9	3	12
		C	2	2	4
	Todos	A	21	13	34
		B	4	4	8
		C	1	3	4
Competencias Ciudadanas	Casa	A	22	4	26
		B	5	1	6
		C	5	2	7
	Sitio	A	17	7	24
		B	5	2	7
		C	7	1	8
	Todos	A	21	3	24
		B	8	1	9
		C	4	2	6

Prueba	Tipo	NCDIF	No Uniforme	Uniforme	Total
Lectura Crítica	Casa	A	20	8	28
		B	5	1	6
		C	4	1	5
	Sitio	A	22	4	26
		B	7	2	9
		C	1	3	4
	Todos	A	19	9	28
		B	4	2	6
		C	4	1	5
Razonamiento Cuantitativo	Casa	A	31	5	36
		B	0	1	1
		C	3	0	3
	Sitio	A	27	7	34
		B	2	1	3
		C	3	0	3
	Todos	A	30	5	35
		B	1	1	2
		C	3	0	3

Examen Saber Pro | Examen Saber TyT | icfes

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