MINISTERIO DE EDUCACIÓN NACIONAL

Comparación de pruebas Saber 3°, 5°, 9° en computador de 2021 y en papel y lápiz de 2022

Subdirección de Estadísticas

Dirección de Evaluación

Agosto 2022

Pruebas Saber 3° 5° 7° 9° | icfes

MINISTERIO DE EDUCACIÓN NACIONAL

Elaboración del documento
Luis Adrián Quintero Sarmiento
John Alexander Calderón Rodríguez
Andrés Ricardo Rodríguez Nagles
Nila Fernanda Amaya Melo

Fotografía portada
https://www.pexels.com/es-es/foto/colegialas-diversas-felices-en-uniforme-sonriendo-junto-a-la-pared-de-ladrillo-5896583/

DIseño y diagramación
Kevin Ostos Peñaloza

ISBN: En trámite

Bogotá D.C., agosto 2022

Comparación de pruebas Saber 3°, 5°, 9° en computador de 2021 y en papel y lápiz de 2022

Pruebas Saber 3° 5° 7° 9°

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**

Pruebas Saber 3° 5° 7° 9°

.10

Introducción

El Instituto Colombiano para la Evaluación de la Educación (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 (artículo 1, Ley 1324 de 2009). De acuerdo con lo anterior, el presente informe se da en el contexto de los contratos interadministrativos número CO1.PCCNTR.2528072 de 2021 y CO1.PCCNTR.3443608 de 2022, suscritos entre el Ministerio de Educación Nacional (MEN) y el Icfes. El objetivo de este documento es presentar los resultados de la comparación de las pruebas Saber 3°, 5° y 9° aplicadas en computador en 2021 y en papel y lápiz en 2022.

Para brindar una contextualización sobre esta prueba, en el año 2018 el MEN, en conjunto con el Icfes, decidieron reestructurar la versión anterior de Saber 3°, 5° y 9° con el fin de actualizar los marcos conceptuales de las pruebas. Por tal razón, en el año 2019 se realizó el pilotaje de la prueba utilizando una muestra de establecimientos educativos del país. En este pilotaje se buscaba evaluar la nueva estructura, así como el comportamiento psicométrico de los ítems en la población de estudiantes.

Con base en los resultados de la prueba piloto del año 2019, se tenía planeado que la aplicación definitiva y comienzo de una nueva línea base se realizara en el segundo semestre de 2020. El diseño de la aplicación se dividió en dos operaciones, una parte censal no controlada y la otra era una muestra representativa a nivel nacional y controlada. Sin embargo, dadas las condiciones de suspensión de clases presenciales en marzo del 2020 a causa de la COVID-19, no fue posible programar la aplicación durante este año y se determinó cancelar las pruebas.

Por otra parte, se identificó que la oferta censal que se había diseñado no era la viable desde el punto de vista financiero y logístico ya que la aplicación censal propuesta solo podía desarrollarse en condiciones de normalidad académica y total presencialidad. Además, estratégicamente, y debido a las alteraciones que estaba atravesando el sector educativo era necesario reevaluar de qué manera podía el país continuar produciendo óptimamente insumos para las decisiones de política educativa.

De esta manera, en el segundo semestre del año 2020 el equipo Icfes elaboró una propuesta técnica en la que se analizó la viabilidad, precisión y confiabilidad de realizar una medición muestral, destacando las ventajas que esta opción tenía en términos de costos, herramientas investigativas y en la disminución de riesgos. Cabe resaltar que esta propuesta incluía la aplicación para estudiantes de grado 7°, con lo cual se incluiría la medición de aprendizajes para estudiantes iniciando el nivel de secundaria. Esto le iba a permitir al país obtener más y mejor información acerca de la calidad educativa en los grados escolares 3°, 5°, 7° y 9°.

La primera fase del proyecto de las pruebas se realizó durante el modelo de alternancias y fue una aplicación muestral controlada de 2021 que se llevó a cabo en formato electrónico (prueba en computador, en adelante PEC), dado que ha sido de interés realizar una transición en el país hacia las pruebas electrónicas, tal como lo han hecho pruebas internacionales estandarizadas de gran renombre como PISA, PIRLS e ICCS, entre otras.

La segunda fase del proyecto consistió en implementar medidas correctivas para el modelo de oferta educativa cien por ciento presencial teniendo en cuenta los problemas estructurales de conectividad y de fuerza mayor que tiene el país, por lo cual se decidió implementar una aplicación bajo un formato de pruebas en papel y lápiz (PPL) en el primer semestre del año 2022. Esto, con el fin de que el país cuente con una evaluación representativa, precisa y confiable para el sector educativo, investigadores, hacedores de política pública y comunidad educativa en general.

Adicionalmente, este plan de continuidad garantizará la posibilidad de adelantar un estudio y análisis comparativo entre la aplicación bajo un modelo de alternancia como fue la modalidad en 2021 y la fase presencial en 2022 en la cual los estudiantes no tienen restricciones de asistencia debido a la pandemia. Este será el primer estudio a nivel Latinoamericano de este tipo y será un referente en evaluación estandarizada y de gran escala para la región.

En el presente estudio el énfasis será en la comparabilidad de las PEC de 2021 y PPL de 2022. Las características de los datos obtenidos a partir de las dos aplicaciones ofrecen

una gran oportunidad para llevar a cabo el estudio de comparabilidad, ya que se tiene información para varios grados escolares, lo cual permite analizar si hay una tendencia presente entre distintos niveles educativos. Además, el diseño de la muestra de 2022 aseguró que se tuviera un número importante de colegios de traslape con la aplicación electrónica llevada a cabo en 2021, lo cual ofrece la posibilidad de analizar la comparabilidad con la misma muestra de sedes y estudiantes traslape en los dos años, dando mayor robustez al estudio de equivalencia entre vehículos de aplicación.

En esta línea, el presente documento describe inicialmente las características que tuvo la aplicación de Saber 3°, 5°, 9° en 2021 y 2022. Posteriormente, se discuten algunos estudios relevantes en la literatura sobre la equivalencia entre formatos PEC y PPL, la metodología utilizada en el presente estudio para comparar los dos vehículos de aplicación, y, finalmente, se presentan los resultados y las principales conclusiones sobre el cambio de formato para estudiantes de educación básica primaria y media en Colombia.

1.1. Aplicación de Saber 3°, 5°, 9° en 2021 y en 2022

El examen Saber 3°, 5°, 9° es una evaluación nacional de carácter externo que se aplica periódicamente a estudiantes de educación básica de todo el país, con el fin de conocer el desarrollo de sus competencias en algunas áreas definidas previamente y en hitos específicos durante el proceso educativo. Este proceso evaluativo es relevante ya que permite identificar fortalezas y oportunidades de mejora en la calidad educativa de los niveles de básica y media. Lo anterior se logra ya que estas pruebas están alineadas con los Estándares Básicos de Competencias establecidos por el MEN (Icfes, 2022a).

La aplicación PEC de las pruebas Saber 3°, 5° y 9° en 2021 se realizó entre los meses de octubre y noviembre, mientras que la aplicación PPL de 2022 se llevó a cabo en abril para los grados 4°, 6° y 10°, es decir que hubo una diferencia de alrededor de cinco meses entre una aplicación y la otra. Dada la forma como se seleccionó la muestra de 2022, se aseguró que se tuvieran 127 sedes en común entre las dos aplicaciones (traslape del 10,2% ya que la muestra en 2022 fue de 1244 sedes), lo cual ofrece una mayor robustez para el estudio de comparabilidad al tener un mismo subconjunto de sedes y estudiantes evaluados bajo los dos formatos. A partir del documento de identidad de los evaluados se pudieron identificar los estudiantes que presentaron la prueba tanto en 2021 como en 2022. Este será el conjunto de datos utilizado en el presente análisis, ya que se facilita la interpretación y extracción de conclusiones al tener los mismos individuos tanto en PEC como en PPL, puesto que, si hay diferencias en los resultados, estas pueden atribuirse al cambio de formato y no a un cambio en la muestra de análisis.

Respecto a las pruebas que fueron aplicadas, en Saber 3° se evaluaron los módulos de Competencias Comunicativas en Lenguaje: Lectura y Matemáticas. Por otra parte, para Saber 5° y 9° se evaluaron estas dos pruebas y, además, Competencias Comunicativas en Lenguaje: Escritura, Competencias Ciudadanas: Pensamiento Ciudadano, Competencias Ciudadanas: Acciones y Actitudes, y, por último, Ciencias Naturales y Educación Ambiental.

Este informe se orienta al análisis de las pruebas cognitivas de opción múltiple con única respuesta, las cuales se califican con un modelo para ítems dicotómicos que permite realizar comparaciones y análisis de forma directa. Por lo anterior, se excluyen de este análisis las pruebas de Competencias Comunicativas en Lenguaje: Escritura y Competencias Ciudadanas: Acciones y Actitudes. Por otro lado, dado que el examen Saber 7° fue aplicado como un piloto[1] en el 2021 y 2022, este grado no se incluye en los análisis de comparabilidad del presente documento.

Respecto al diseño de las pruebas, este fue similar en 2021 y 2022, exceptuando el formato de aplicación. En la **Tabla 1** se presenta el número de formas/cuadernillos empleados para la evaluación de cada prueba en las dos aplicaciones y el número de ítems evaluados. Las pruebas de Lectura y Matemáticas de grado 3° se evaluaron utilizando seis formas y 90 ítems, mientras

1 El piloto se refiere a la fase previa durante el desarrollo de un instrumento definitivo que consiste en aplicar un grupo de ítems a una muestra para recoger información relacionada con la calidad de estos ítems. A partir de esta información se identifican oportunidades de mejora en ellos y se toman decisiones que permitan mejorar la prueba

que en 5° se emplearon doce formas para Lectura (120 ítems) y Matemáticas (108 ítems) y seis formas (90 ítems) para Ciencias Naturales y Pensamiento Ciudadano. Cabe resaltar que las formas tienen ítems en común entre ellas para poder asegurar que los resultados de los estudiantes que presentan distintas formas se encuentren en una misma escala. Por último, en 9° se emplearon doce formas (126 ítems) para Lectura y Matemáticas y seis formas (108 ítems) para Ciencias Naturales y Pensamiento Ciudadano. En 5° y 9° se tienen 12 formas para Lectura y Matemáticas debido al componente de trayectorias escolares incluido para estas dos pruebas, en el cual se evalúan los estudiantes incluyendo ítems de otros grados para así tener una escala de calificación comparable entre grados.

El presente estudio tiene la ventaja de contar con un número alto de preguntas en cada prueba gracias a que se tienen múltiples formas, lo cual enriquece el análisis ya que se tienen ítems con características más diversas que pueden ayudar a tener una visión más amplia de la comparabilidad entre PEC y PPL. Por lo general, los estudios que comparan los dos formatos tienen solo una forma para cada prueba y, por lo tanto, un número reducido de ítems.

Todos los estudiantes de 3°, 5° y 9° fueron evaluados en Lectura y Matemáticas, mientras que cada estudiante de 5° y 9° fue evaluado además en dos de las otras cuatro áreas restantes. En otras palabras, mientras un evaluado de estos grados presentaba Acciones y actitudes y Escritura, otro evaluado presentaba Ciencias naturales y Pensamiento ciudadano. Lo anterior, para un total de cuatro pruebas por estudiante en los grados 5° y 9°. Como se puede observar en la **Tabla 1**, los estudiantes de 3° y 5° presentaron formas de 30 ítems en cada prueba, mientras que los estudiantes de 9° presentaron formas de 36 ítems.

Recordemos que no todos los niños de la aplicación PEC hacen parte de la muestra para la aplicación de PPL. Por tanto, en el presente análisis, se tuvieron en cuenta exclusivamente los estudiantes que presentaron la prueba tanto en 2021 como en 2022 para tener la misma población en PEC y en PPL. A través del documento de identificación de los evaluados se pudo establecer el

Tabla 1. Cantidad de formas e ítems empleados en la evaluación de cada una de las pruebas en Saber 3°, 5°, 9° 2021 y 2022.

Prueba	Formas	Ítems por forma	Ítems en total
Tercero			
Lectura	6	30	90
Matemáticas	6	30	90
Quinto			
Ciencias Naturales y Educación Ambiental	6	30	90
Pensamiento Ciudadano	6	30	90
Lectura	12	30	120
Matemáticas	12	30	108
Noveno			
Ciencias Naturales y Educación Ambiental	6	36	108
Pensamiento Ciudadano	6	36	108
Lectura	12	36	126
Matemáticas	12	36	126

conjunto de estudiantes en común para las dos muestras. Además, para estos análisis se utilizaron los evaluados sin reporte de ningún tipo de discapacidad, que hubieran respondido como mínimo el 50% de los ítems en cada prueba y que no fueran sospechosos de copia.

En la **Tabla 2** se presenta el número de estudiantes total en el presente análisis desagregando por grado, sexo, zona y sector. Como se puede observar, se tienen alrededor de 5.000 estudiantes en cada grado y la muestra es balanceada por sexo. Al mirar las cantidades por zona y sector, se encuentra que alrededor del 15% de los estudiantes por grado estudian en zona rural y alrededor del 25% en establecimientos no oficiales, reflejando en cierta medida la realidad nacional en cuanto a las características de los establecimientos educativos.

Tabla 2. Cantidad de estudiantes en el estudio que presentaron la prueba en formato PEC en 2021 y en PPL en el 2022, desagregando por sexo, zona y sector.

Categorías	Tercero		Quinto		Noveno	
	Cantidad	(%)	Cantidad	(%)	Cantidad	(%)
Sexo						
Niña/Mujer	2002	47%	2372	49%	2970	53%
Niño/Hombre	2246	53%	2437	51%	2643	47%
Zona						
Rural	591	14%	768	16%	762	14%
Urbana	3657	86%	4041	84%	4851	86%
Sector						
No oficial	1122	26%	1091	23%	1294	23%
Oficial	3126	74%	3718	77%	4319	77%
Total	**4248**	**100%**	**4809**	**100%**	**5613**	**100%**

1.2. Antecedentes

Las tecnologías de la información y comunicación desempeñan un papel cada vez más preponderante en la sociedad actual, no solo en la vida escolar de los estudiantes del país, sino que en general en el consumo de información y en la comunicación digital, las cuales se han vuelto determinantes para interactuar social y laboralmente. Por lo tanto, la aplicación de pruebas electrónicas resulta un componente importante para el sistema educativo y para el desarrollo de personas integrales y preparadas para participar con oportunidades de los mercados laborales y de la democratización de la información.

Por lo tanto, el Icfes ha venido desarrollando pruebas en formatos electrónicos en los últimos años, las cuales además tienen múltiples ventajas como reducir los costos por impresión y transporte del material. Siguiendo esta línea, las pruebas Saber 3°, 5°, 9° no han sido ajenas a este proceso. A continuación, se reporta el historial de pruebas aplicadas por el Icfes en formato electrónico, así como los estudios previos que se han realizado sobre comparabilidad entre pruebas en PEC y en PPL.

1.2.1. Pruebas Icfes aplicadas en formatos electrónicos

Para realizar la implementación de aplicaciones en formato electrónico, el Instituto creó una herramienta denominada 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 través de los ítems y devolverse a preguntas que ya se hayan resuelto previamente para la prueba que esté respondiendo (Icfes, 2020).

A través de PLEXI, el Icfes ha desarrollado diversas pruebas en formato PEC. De acuerdo con el Informe de Gestión de 2019 entregado por el Instituto, durante ese año se realizaron 15 pruebas electrónicas por medio de PLEXI (Ver **Tabla 3**). 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 y Policía Nacional extemporánea, estas se aplicaron también por medio de PLEXI, utilizando los equipos de cómputo que se encontraban en los sitios de aplicación. El total de evaluados por medio del aplicativo fue de 4.397.

En el año 2020, a raíz de la pandemia, las pruebas Saber Pro y Saber TyT se realizaron como PEC, y se continúan aplicando en este formato hasta el día de hoy. Por otro

Tabla 3. Pruebas desarrolladas en formato electrónico durante el 2019.

Prueba	Formas
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	567.360

Nota: Tomado de Icfes (2019).

lado, en el 2020 el Icfes desarrolló Evaluar para Avanzar que es una herramienta de uso voluntario para apoyar y acompañar los procesos de enseñanza de los y las docentes de grados tercero a once. En el año 2020 se contó con la participación de 338.395 estudiantes que finalizaron la aplicación electrónica en modalidad online u offline de alguna de las ocho pruebas disponibles[2], mientras que en el 2021 se tuvieron 1'636.413 estudiantes y en el 2022 esta cantidad aumentó a 3'202.636, lo cual demuestra la gran aceptación que ha tenido la herramienta por parte de la comunidad educativa.

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, 2019). Esta metodología brinda mayor flexibilidad y requiere un menor número de preguntas para evaluar a los estudiantes con una precisión adecuada.

El avance en la integración de la metodología CAT en PLEXI dio como resultado que el Instituto pudiera ofrecer a la comunidad de estudiantes la herramienta El Icfes tiene un preicfes, una herramienta virtual a la que pueden acceder gratuitamente los estudiantes que desean familiarizarse con las preguntas de los exámenes de Estado. Esto es importante ya que, para algunos estudiantes, las pruebas pueden generar estrés o ansiedad (Crooks, 2004; Heissel, 2014) y estas variables pueden afectar los resultados de los estudiantes generando rendimientos bajos (OCDE, 2013), de manera que es importante ofrecerles a los estudiantes la posibilidad de tener un acercamiento previo con las pruebas.

1.2.2. Comparabilidad de pruebas Icfes en PEC y PPL

Cuando se desea aplicar una prueba como PEC y PPL simultáneamente, 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 y algunos se han realizado bajo el contexto de las pruebas Saber 3°, 5°, 9°. Por ejemplo, el piloto de este examen en 2019 fue aplicado en papel y lápiz y se seleccionó una submuestra (a conveniencia) de colegios que contaban con equipos de cómputo y conexión a internet, de manera que se les aplicó la prueba de manera electrónica. Quintero et al. (2022) analizaron la comparabilidad del examen Saber 3°, 5°, 9° en los formatos PEC y PPL utilizando métodos cuasiexperimentales y encontraron 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 tomó a conveniencia de acuerdo con la disponibilidad de computadores e internet, se debe tener cuidado con el alcance de las conclusiones.

Además, Quintero et al. (2022) encontraron que el número de ítems con funcionamiento diferencial entre los dos formatos disminuye a medida que aumenta el grado de escolaridad. Al revisar el texto y la presentación de los ítems en la plataforma, los autores encontraron que un elemento determinante para el funcionamiento diferencial es que en las preguntas largas los estudiantes deben desplazarse a través de la pantalla hacia abajo (scrolling) para leer las opciones de respuesta. Como consecuencia, los evaluados de habilidades bajas tienden a seleccionar las opciones de respuesta en la parte superior (A y B) que aparecen primero que las otras opciones al desplazarse verticalmente en la pantalla.

En el año 2020, en medio de la pandemia, se aplicaron las pruebas Saber Pro y Saber TyT en PEC y se analizó su comparabilidad con el formato PPL aplicado desde el segundo semestre de 2016, siendo esta la línea base, según la Resolución 126 de 2016. Los resultados indicaron que se puede asumir equivalencia entre los dos vehículos de aplicación. Estos resultados concuerdan con otros estudios donde se ha encontrado que, a mayor

2 Las ocho pruebas que se encuentran disponibles para el proyecto de Evaluar para Avanzar son: Matemáticas, Competencias Comunicativas en Lenguaje: Lectura, Lectura Crítica, Ciencias Naturales y Educación Ambiental, Ciencias Naturales, Competencias Ciudadanas: Pensamiento Ciudadano, Sociales y Ciudadanas, Inglés, y por último, Cuestionarios Auxiliares. (Icfes, s.f.)

escolaridad de los evaluados, hay mayor comparabilidad entre los dos formatos (Choi & Tinkler, 2002). Por lo tanto, se decidió seguir aplicando estas dos pruebas para estudiantes de educación superior en formato electrónico.

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 PEC y PPL a estudiantes de sexto de secundaria de la República Dominicana (Icfes, 2022b). El estudio arrojó diferencias significativas entre los promedios en los dos formatos para las pruebas de Matemática y Lengua Española. La diferencia entre los promedios resultó pequeña para Matemáticas, con puntajes más altos para los estudiantes que presentaron la prueba en computador, y mediana para Lengua Española, a favor de la aplicación en PPL. Estos resultados sugieren que no hay comparabilidad directa al evaluar los estudiantes en los dos formatos, por lo cual es importante entender más en detalle las razones de las diferencias, especialmente si el Icfes desea hacer una transición hacia las PEC y espera que haya una equivalencia con los resultados obtenidos históricamente en PPL.

El presente estudio permite realizar un análisis más completo de comparabilidad en el contexto colombiano, y tiene múltiples ventajas en comparación con el estudio realizado con el piloto de Saber 3°, 5°, 9° en 2019. La principal diferencia radica en que la muestra del presente estudio no fue seleccionada a conveniencia, sino de manera aleatoria, ya que se transportaban los equipos a las escuelas para que los estudiantes pudieran aplicar la prueba en PEC. Además, en el presente análisis se tiene un mismo subconjunto de estudiantes que presentaron la prueba en los dos formatos, lo cual permite realizar comparaciones sin riesgo de sesgos por tener poblaciones distintas en cada vehículo de aplicación. Esto permite a su vez obtener conclusiones más robustas en el estudio.

02.

Metodología

El estudio de equivalencia entre formatos involucra una serie de pasos para comparar el comportamiento de los ítems en los dos vehículos de aplicación y los puntajes obtenidos por los estudiantes. A continuación, se detallan los análisis realizados desde la teoría clásica del test y la teoría de respuesta al ítem.

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, es decir en la habilidad o competencia 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 el examen en PEC y en PPL, 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 PPL y en PEC.

Múltiples esfuerzos se han realizado para estudiar la equivalencia de formatos. Algunos estudios han encontrado que los puntajes en PEC y PPL son comparables (Bridgeman et al., 2003; Poggio et al., 2005), mientras que otros han concluido que los puntajes de los dos vehículos de aplicación no son equivalentes (Carlbring et al., 2007; McCoy et al., 2004; Pommerich, 2004). Choi & Tinkler (2002) concluyeron que los ítems en PEC son más difíciles que en PPL y que los efectos de formato son mayores para estudiantes de tercer grado que para estudiantes de décimo. Bennet et al. (2008) encontró que el puntaje promedio para estudiantes de octavo era significativamente menor en PEC comparado con PPL, pero la diferencia es muy pequeña en términos del tamaño del efecto. Way et al. (2008) analizó múltiples estudios de la prueba estadounidense K-12 y concluyó que distintas tendencias emergen dependiendo de la prueba y el grado evaluado, pero había efectos pequeños o insignificantes en la mayoría de los casos. Piaw (2011) comparó PEC y PPL para dos pruebas psicológicas y concluyó que PEC era más interpretable en términos de validez interna y externa, además de reducir el tiempo de aplicación y aumentar la motivación de los evaluados. Sin embargo, los estudiantes en PEC no obtuvieron puntajes más altos que en PPL. Por último, Jerrim (2016) analizó los puntajes de PISA 2012 de 32 países/economías y encontró puntajes más bajos para las PEC en 11 de ellos, y puntajes más altos en PEC en 13 países, pero la magnitud de las diferencias fue pequeña en la mayoría de los casos.

En la literatura, la equivalencia entre los dos formatos se evalúa principalmente 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 PEC y en PPL.

2.2. Análisis de ítem

Con el fin de estudiar las propiedades psicométricas de las pruebas en PPL y en PEC, se realizó un análisis univariado y un análisis desde la teoría clásica del test (TCT) y la teoría de respuesta al ítem (TRI). Por su parte, el análisis univariado se realizó con el objetivo de analizar si hay opciones de respuesta con porcentajes de respuesta asociados inferiores al 5% o superiores al 95% en toda la prueba y verificar si las respuestas de los estudiantes se concentraron en la clave de los ítems.

Además, se identificaron aquellos ítems que presentaron comportamientos atípicos tales como altos porcentajes de omisión, es decir, con más del 5 % de no respuesta. Para la prueba en PPL se revisaron además los ítems con más del 5% de evaluados con multimarca. Esto no

aplica para PEC ya que la plataforma permite seleccionar sólo una respuesta en cada pregunta. El porcentaje de omisiones y multimarca puede ser indicador de algunas dificultades o actitudes de los estudiantes en las diferentes formas conformadas en cada prueba, ya que pueden estar evidenciando preguntas que no se alcanzaron a responder por falta de tiempo, aplazamiento de la respuesta, desconocimiento de la respuesta correcta, entre otros aspectos. De igual forma, el porcentaje de multimarca puede deberse a una falta de comprensión de las instrucciones brindadas, dudas frente a la selección de la respuesta correcta, confusiones al momento de marcar, entre otras (Icfes, 2022b). Los ítems identificados con estas características se señalaron para a un proceso de revisión más detallado con el fin de determinar las posibles razones para tales comportamientos.

En lo que tiene que ver con la TCT se realizaron análisis para establecer las bondades del instrumento. En teoría, si el instrumento es válido, el puntaje de cada ítem debería correlacionarse positivamente con los puntajes globales obtenidos en el instrumento. Para evaluar esa correlación ítem-prueba se procedió a examinar el índice de correlación punto biserial entre la respuesta al ítem (que es una variable discreta con dos posibles valores, esto es, 1 si la respuesta es correcta y 0 en caso contrario) y el puntaje global (que es una variable continua porque resulta de la suma de aciertos del evaluado). El procedimiento supone examinar las relaciones sistemáticas entre la variable discreta y la continua. Aquellos ítems que no se relacionen con el resto de la prueba (cercano a cero o negativo) fueron revisados con un mayor nivel de detalle con el equipo de expertos de cada prueba junto con los ítems que presentaron alertas desde el análisis univariado en cuanto a porcentajes altos de omisión y multimarca.

Luego de revisar los ítems, se seleccionan aquellos que definitivamente tienen un comportamiento psicométrico atípico y se eliminan del proceso de calificación que procede a partir de la TRI. Con las estimaciones obtenidas del modelo se pueden identificar también ítems con comportamientos extremos desde la TRI en cuanto a sus parámetros de dificultad y discriminación como se explica en la siguiente sección. Estos ítems también se revisan para una posible eliminación del modelo de calificación de acuerdo con los puntos de corte definidos en la literatura. A continuación, se discuten las metodologías utilizadas para llevar a cabo dicha calificación.

2.3. Modelo de calificación

Las pruebas Saber se califican con base en modelos de 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 3°, 5°, 9° se califica a partir del modelo TRI de dos parámetros (2PL), de manera que, para cada ítem i, se estima un parámetro de dificultad (b_i) y un parámetro de discriminación (a_i). Bajo este modelo, la probabilidad de que el estudiante j conteste correctamente al ítem i se define como:

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

Donde θ_j es la habilidad del estudiante j, a_i determina la pendiente máxima de la curva característica de cada ítem, el parámetro b_i es el punto de la escala de habilidad donde la probabilidad de responder correctamente al ítem i es igual a 0,5 y el parámetro $D=1,702$ es un factor de escala. 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 para determinar sus características psicométricas. Como parte del análisis de ítem descrito anteriormente, desde la TRI se revisan los ítems cuyo parámetro de dificultad esté por fuera del rango [-3, 3] y con valores de discriminación menores a 0,25.

2.4. Confiabilidad marginal

Una característica importante de la prueba que se puede comparar para analizar la equivalencia entre formatos es el valor de la confiabilidad a partir de las versiones en PEC y PPL. Sin embargo, debido a que en el presente estudio se utilizaron múltiples formas/cuadernillos en cada prueba, no se puede calcular directamente la

confiabilidad TCT a través del coeficiente KR-20 (Kuder y Richardson, 1937), el cual requiere que todos los estudiantes hayan respondido las mismas preguntas. Por lo tanto, para comparar la confiabilidad de las pruebas en los dos formatos de aplicación, se emplea la confiabilidad marginal, la cual se calcula a partir del modelo de TRI.

Dado que en los modelos de TRI el error de estimación se puede expresar en términos de la habilidad a través de la función de información (Green et al., 1984), es posible realizar una estimación de la confiabilidad como medida resumen de la precisión del test a partir de la confiabilidad marginal, definida por la siguiente expresión:

$$\rho = 1 - \left(\frac{ee^2_M}{\sigma^2_\theta} \right)$$

donde $ee^2_M = \int ee^2(\theta) d(\theta) d\theta / \int d(\theta) d\theta$ es la integral del error cuadrático de medición y σ^2_θ es la varianza de los puntajes. La confiabilidad marginal toma valores entre 0 y 1, tal que valores cercanos a 1 indican que la prueba tiene una precisión adecuada.

2.5. 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/cuadernillos para la prueba (Kolen y Brennan, 2014). Esto es de gran importancia en el presente estudio, ya que es necesario asegurar que los puntajes de estudiantes que presentaron distintas formas/cuadernillos en la prueba se encuentren en la misma escala y sean comparables.

La equiparación también es útil para el análisis de equivalencia entre formatos, ya que esta herramienta permite llevar los puntajes de los dos vehículos de aplicación a una misma escala y realizar comparaciones. Existen varias metodologías para realizar el proceso de equiparación, especialmente 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).

Para el presente estudio se utilizó el método de calibración conjunta de ítems (Bock y Zimowski, 1997) para garantizar la comparabilidad de los puntajes entre formas/cuadernillos dentro de cada vehículo de aplicación. Esta metodología asume que los parámetros de dificultad y discriminación son iguales entre formas, y se estiman los parámetros de manera simultánea para las distintas formas que comparten un mismo ítem. El método de calibración conjunta también se puede utilizar para realizar la equiparación entre PEC y PPL, ya que, al estimar conjuntamente los parámetros de los ítems en los dos vehículos de aplicación, los puntajes resultantes se encuentran en la misma escala.

Sin embargo, antes de llevar a cabo esta equiparación, es necesario determinar cuáles ítems se comportan de manera similar en los dos formatos. Esto permite asumir que los parámetros de dichos ítems en el modelo de calificación son los mismos en los dos vehículos de aplicación para así realizar su estimación conjunta. 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. Para los ítems con DIF, se permite que sus parámetros sean distintos en PEC y PPL en el proceso de estimación, ya que no se puede asumir igualdad para ellos entre los dos formatos.

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

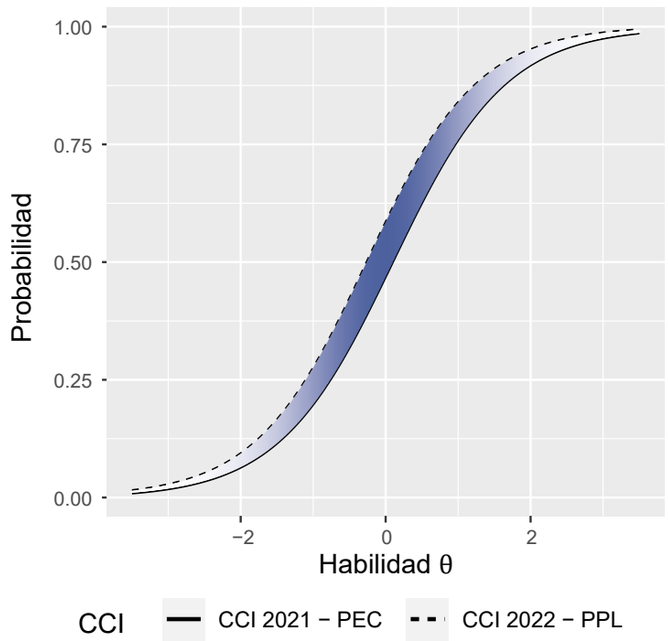
en los dos formatos se puedan equiparar correctamente. Usualmente la literatura sugiere un número mínimo de 15 ítems sin DIF, o, en pruebas largas, al menos el 20% del número total de ítems en la prueba (Kang y Petersen, 2012). Dado lo anterior, a continuación, se presentan generalidades del análisis de DIF.

2.6. 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 (ver Figura 1). La CCI es una función que depende de los parámetros de dificultad y discriminación de cada ítem y relaciona la habilidad de los estudiantes (eje horizontal) con la probabilidad de responder correctamente al ítem (eje vertical). Si hay diferencia entre la CCI de PEC y PPL, esto puede indicar sesgo en la medición que, para este contexto, se refiere a favorecer a un formato de aplicación sobre otro en la evaluación. En el presente estudio, los dos grupos que se analizarán para determinar la invarianza entre formatos corresponden a: 1) la población relacionada con la evaluación en PEC en 2021, y 2) la población de la aplicación de 2022 en PPL.

A partir de este análisis, los ítems sin DIF entre PEC y PPL serán fijados o anclados en la calificación para los dos formatos, mientras que ítems con DIF deben ser calibrados de forma independiente para cada formato con el fin de incluirse en la medición de cada grupo. La metodología usada en los análisis de DIF se basa en la propuesta de Raju (1988), que consiste en estudiar las diferencias entre las CCI para los dos grupos de comparación, es decir, cuantificar la magnitud de las diferencias a lo largo del rasgo latente para identificar cuándo esta es grande, lo cual implica que el ítem presenta DIF (ver **Figura 1**).

Figura 1. Ejemplo de la CCI en PEC y PPL para un ítem de la prueba.



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).

Para la implementación de la metodología de DIF, se deben 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 en PEC, y el grupo de referencia es el grupo de estudiantes que presentaron la prueba en formato PPL. 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 CCI 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. La metodología del tamaño 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. A partir de estos resultados, se identifican los ítems con categoría A de DIF para anclar entre formatos y tener los puntajes de PEC y PPL en una misma escala. Los ítems en categorías B y C se estiman de manera independiente en cada uno de los formatos.

2.7. Calificación con teoría de respuesta al ítem

Para la obtención de puntajes en la calificación de cada prueba, se ajusta un modelo TRI de dos parámetros (2PL) a las cadenas de respuestas de los evaluados. Teniendo en cuenta que las pruebas Saber 3°, 5°, 7° y 9° son de carácter diagnóstico, se tuvo un tratamiento especial para los ítems que los estudiantes no alcanzaron a responder debido a la longitud y al tiempo disponible para la prueba, tal como lo hace PISA bajo la metodología de no alcanzados (Pisa, 2016). En esta metodología, los ítems no abordados consecutivamente al final de la prueba no se penalizan tomándolos como respuestas incorrectas (como se hace para otras pruebas tales como Saber 11) sino que se excluyen en el momento de la calificación poniendo su respuesta como NA (información faltante), excepto por el primer ítem de la serie de ítems no abordados, el cual se incluye como una respuesta incorrecta.

Por otro lado, como se mencionó anteriormente, para tener puntajes comparables bajo los dos formatos de aplicación, se utilizó el método de calibración conjunta de ítems (Bock y Zimowski, 1997). En este método de equiparación se estiman conjuntamente los parámetros de dificultad y discriminación de los ítems sin DIF para que sean iguales en los dos formatos. Esto es posible ya que se aplicaron las mismas formas y los mismos ítems en 2021 y 2022. Para los ítems que reportaron DIF entre formatos, los parámetros se estiman de forma independiente en cada vehículo de aplicación, lo cual permite que los parámetros de dificultad y discriminación sean distintos entre PEC y PPL para esos ítems.

Como resultado de la estimación del modelo TRI, se obtienen los puntajes de los estudiantes en una escala con media igual a cero y desviación estándar igual a 1. Por lo tanto, para facilitar la interpretación de los resultados, se transforman los puntajes a una escala con media igual a 300 y desviación estándar de 50 puntos. Esta escala es distinta a la escala de publicación que tiene media de 400 y desviación estándar igual a 80. Esto se hace para evitar que se realicen comparaciones directas entre los resultados de los dos informes, los cuales tienen objetivos distintos.

Dado que los puntajes TRI obtenidos se encuentran en una misma escala para PPL y PPC a partir de los ítems que no presentan DIF, es posible comparar el desempeño de los estudiantes en los dos formatos mediante el cálculo de promedios. De igual manera, se puede analizar si la diferencia entre formatos es más pronunciada para algunos grupos de interés como son los estudiantes en zona rural, zona urbana, establecimientos educativos privados, establecimientos educativos públicos, etc. Esto permite analizar si utilizar un entorno de evaluación, llámese electrónico o lápiz y papel, puede castigar o privilegiar más a algunos grupos específicos.

Para la comparación de promedios, se puede analizar si hay diferencias significativas entre PEC y PPL. 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 de los promedios y se calcula de la siguiente manera:

$$TE = \frac{\mu_{PPL} - \mu_{PEC}}{\sigma_{PPL}}$$

Donde μ_{PPL} y μ_{PEC} son, respectivamente, el promedio de los puntajes en PPL y en PEC, mientras que σ_{PPL} es la desviación estándar de los puntajes en PPL. A partir del valor del TE se puede determinar si la diferencia es despreciable (valores menores a 0,1), pequeña (entre 0,1 y 0,3), mediana (entre 0,3 y 0,6) o grande (mayores a 0,6), según los puntos de corte definidos en Cohen (1988). efectiva por ítem y por grado en cada una de las regiones

Pruebas Saber 3° 5° 7° 9°

03.

Resultados

A continuación, se presentan los resultados obtenidos para la comparabilidad de formatos, teniendo en cuenta los análisis propuestos en el apartado de metodología. Inicialmente, se realiza un análisis de DIF para determinar el número de ítems que tienen un comportamiento diferencial al ser aplicados en PEC y en PPL ya que, de esta manera, se puede evaluar la posibilidad de tener una misma escala para los dos formatos, realizando un proceso de equiparación en caso de que haya un número adecuado de ítems sin DIF. Luego, se comparan los puntajes obtenidos en los dos formatos para cada estudiante a través del cálculo de correlaciones. Por último, se comparan los puntajes en los dos formatos en términos de promedios, con el fin de determinar si hay diferencias entre vehículos de aplicación.

3.1. Análisis de DIF

En la **Tabla 4** se presenta el número total de ítems que se tuvieron en cuenta para la calificación después de eliminar algunos de ellos con base en el análisis de ítem, como se explicó en la sección de metodología. Se tuvieron entre 77 y 118 ítems al tener en cuenta todas las formas para el análisis de cada una de las pruebas. Asimismo, se presenta el valor del DIF para estos ítems de acuerdo con la estadística NcDIF. Como se puede observar, en todos los casos el porcentaje de ítems con un efecto grande de DIF (C) es menor al 20%. Este porcentaje corresponde a los ítems que posiblemente conllevan a un proceso cognitivo distinto cuando los estudiantes los responden en PEC y en PPL, ya sea porque contienen gráficos, por ejemplo, o porque contienen textos largos (que implican scrolling), entre otras posibles causas. El porcentaje de ítems con tamaño de DIF moderado (B) es relativamente bajo también, de modo que la mayoría de los ítems tuvieron un tamaño de DIF insignificante (A).

Al realizar el análisis por grado, se observa que hay una proporción baja de ítems en la categoría C para noveno (entre 2 y 8%), lo cual implica que se presentan pocas diferencias, a nivel de ítems entre los formatos para este grado. Para 3° y 5° se identifica que entre el 4 y 19% de los ítems presentan DIF en categoría C, es decir que se encuentra una mayor proporción de ítems con comportamiento diferencial en comparación con 9°. En cuanto a las pruebas, se observa que Ciencias naturales y Competencias ciudadanas tienen una menor proporción de ítems con DIF insignificante (A), es decir que tienden a tener una mayor cantidad de ítems con comportamiento diferencial (categorías B y C) en comparación a Lectura y Matemáticas. El hecho de que la mayoría de los ítems presenten categoría A de DIF es un indicio de que puede haber comparabilidad entre los dos formatos, ya que las propiedades psicométricas de los ítems son similares independientemente del vehículo de aplicación.

Sin embargo, se debe analizar si hay comparabilidad a nivel de puntajes como se hace a continuación a partir del cálculo de correlaciones y promedios para los vehículos de aplicación. Por otro lado, es importante considerar que aparte de tener equivalencia entre formatos, el tránsito hacia PEC en el país implica retos en términos operativos para asegurar que todos los estudiantes puedan resolver las pruebas en computador.

Tabla 4. Cantidad (porcentaje) de ítems en cada categoría de DIF de acuerdo con su tamaño del efecto.

Prueba	# total de ítems	A	B	C
Tercero				
Lectura	77	64 (83%)	10 (13%)	3 (4%)
Matemáticas	85	53 (62%)	19 (22%)	13 (15%)
Quinto				
Lectura	109	92 (84%)	9 (8%)	8 (7%)
Matemáticas	101	79 (78%)	12 (12%)	10 (10%)
Ciencias N.	84	51 (61%)	17 (20%)	16 (19%)
Competencias C.	86	58 (67%)	18 (21%)	10 (12%)
Noveno				
Lectura	118	108 (92%)	7 (6%)	3 (3%)
Matemáticas	109	100 (92%)	7 (6%)	2 (3%)
Ciencias N.	104	79 (76%)	17 (16%)	8 (8%)
Competencias C.	104	89 (86%)	7 (7%)	8 (8%)

Nota: En algunos casos la suma de los porcentajes no es igual al 100% como resultado de la aproximación

3.2.Confiabilidad marginal

Como se explicó anteriormente, para la calificación de los estudiantes se realizó una calibración conjunta asumiendo igualdad en los parámetros de los ítems con categoría de DIF insignificante (A) para PEC y PPL, de manera que se tenga una misma escala para los dos formatos. El porcentaje de ítems con DIF insignificante fue igual o mayor al 61% en cada prueba-grado, de manera que se tuvo una cantidad adecuada de ítems para anclar entre vehículos de aplicación. Para los ítems con DIF en categorías B y C se permitió que tengan parámetros distintos en PEC y PPL, ya que no se podía asumir su equivalencia entre formatos de acuerdo con los resultados del análisis de DIF.

En la **Tabla 5** se reporta la confiabilidad marginal al realizar la calificación para cada prueba y, como se puede observar, en todos los casos se tiene una confiabilidad alta, lo cual indica que los evaluados se están midiendo de manera adecuada en cuanto al error de medición. Por otro lado, la confiabilidad de cada prueba es bastante similar en los dos formatos, lo cual es un indicador inicial de que puede haber comparabilidad entre vehículos de aplicación. La diferencia entre la confiabilidad de los dos formatos está entre 0,1 o 0,3 en todos los casos, a favor de PPL en todas las pruebas, excepto por Lectura en grado 9°.

Tabla 5. Confiabilidad marginal en cada formato, diferencia entre las confiabilidades y la correlación de los puntajes de cada estudiante en PEC y PPL.

Prueba	Conf. Mar. PEC	Conf. Mar. PPL	Difer. Conf.	Correlación
Tercero				
Lectura	0,84	0,86	0,03	0,59
Matemáticas	0,84	0,86	0,02	0,56
Quinto				
Lectura	0,85	0,86	0,01	0,64
Matemáticas	0,83	0,85	0,02	0,63
Ciencias N.	0,87	0,88	0,01	0,63
Competencias C.	0,81	0,84	0,03	0,56
Noveno				
Lectura	0,85	0,84	-0,01	0,71
Matemáticas	0,82	0,84	0,02	0,74
Ciencias N.	0,87	0,88	0,01	0,69
Competencias C.	0,87	0,89	0,02	0,69

3.3.Comparación de puntajes

Utilizando el hecho de que cada estudiante presentó la prueba en los dos formatos, se puede calcular la correlación del puntaje obtenido por los evaluados bajo los dos vehículos de aplicación. Si el vehículo de aplicación no tuviera un efecto importante, se esperaría una correlación cercana a 1, ya que estudiantes con un puntaje alto en PEC tendrían un puntaje alto en PPL. Como se puede observar en la Tabla 5, las correlaciones tienden a ser más altas cuando aumenta el grado de escolaridad, ya que se tienen valores de correlación entre 0,56 y 0,59 para grado 3°, entre 0,56 y 0,64 para grado 5°, y entre 0,69 y 0,74 para grado 9°.

Esto indica que, especialmente para grado noveno, estudiantes que obtuvieron un puntaje alto en la prueba en formato PEC en 2021, tienden a tener un puntaje alto en formato PPL en 2022. Lo mismo sucede para grados 3° y 5°, pero la correlación es menos fuerte, lo cual indicaría que el formato puede tener un mayor impacto en grados menores. Además de analizar las correlaciones, es importante revisar los puntajes promedio obtenidos en los dos vehículos de aplicación para realizar conclusiones con respecto a la comparabilidad entre formatos.

3.4. Comparación de promedios

La Tabla 6 reporta los promedios de los puntajes en cada formato, la diferencia entre los dos promedios señalando los casos en que la diferencia es significativa al 5% (*) y adicionalmente, el tamaño del efecto (TE) con su clasificación (D - despreciable si es menor a 0,1; P - pequeña si está entre 0,1 y 0,3; M - mediana si se encuentra entre 0,3 y 0,6; G - grande si es mayor a 0,6). Lo ideal para concluir que no hay ninguna diferencia entre el puntaje promedio en los dos formatos sería encontrar un TE despreciable. En caso de encontrar que el TE es pequeño, mediano o grande, se concluiría que hay diferencias importantes entre los puntajes promedio, especialmente en los últimos dos casos.

Como se puede observar en la **Tabla 6**, en grado 3° hay una diferencia despreciable para Matemáticas y pequeña para Lectura, en la cual los estudiantes en PPL obtuvieron un promedio mayor que en PEC. En el caso de 5°, la diferencia es despreciable para todas las pruebas, excepto por Lectura, en donde la diferencia es pequeña a favor de PPL. En el grado 9° la diferencia es pequeña para Matemáticas, Ciencias Naturales y Competencias Ciudadanas, y es mediana para Lectura, a favor de PPL en todos los casos.

Estos resultados no están alineados con lo que indica la literatura en cuanto a que estudiantes en grados más altos de escolaridad tienden a tener un menor efecto del formato de aplicación (Choi & Tinkler, 2002). Esto podría explicarse en parte porque los textos de las preguntas que se utilizan en grado 9° son en general más largos, por lo cual los estudiantes deben desplazarse a través de la pantalla para resolver los ítems, de manera que se reduce la probabilidad de responder correctamente el examen en PEC. Esto también puede explicar que en 3° y 5° haya una diferencia para Lectura, cuyas preguntas tienden a ser más largas comparadas con las de las otras pruebas. Por lo tanto, al tener un número alto de preguntas de mayor longitud en las pruebas de 9°, hay un mayor impacto en los puntajes causado por el vehículo de aplicación.

El hecho de que haya un impacto de formato en 9° no es contradictorio con que haya una correlación más alta para este grado como se encontró en la sección anterior. Lo que esto quiere decir es que estudiantes con puntajes altos en PPL tienden a tener puntajes altos en PEC, pero siempre con una diferencia de alrededor de 15,04 puntos entre los dos formatos en Lectura, por ejemplo. En 3° y 5° la correlación es un poco menor, así que los estudiantes son un poco menos consistentes en cuanto a sus puntajes al aplicarles la prueba en los dos formatos.

Tabla 6. Puntaje promedio de los estudiantes en cada formato, la diferencia entre los promedios y el tamaño del efecto (TE) de acuerdo con el d de Cohen.

Prueba	N estudiantes	Promedio PEC	Promedio PPL	Diferencia	TE
Tercero					
Lectura	4.038	300,00	311,45	11,45*	0,23(P)
Matemáticas	4.052	300,00	302,44	2,44*	0,05(D)
Quinto					
Lectura	4.730	300,00	309,60	9,60*	0,19(P)
Matemáticas	4.727	300,00	304,56	4,56*	0,09(D)
Ciencias N.	1.125	300,00	300,86	0,86	0,02(D)
Competencias C.	1.143	300,00	298,63	-1,37	-0,03(D)
Noveno					
Lectura	5.566	300,00	315,04	15,04*	0,30(M)
Matemáticas	5.553	300,00	310,75	10,75*	0,21(P)
Ciencias N.	1.366	300,00	308,10	8,10*	0,16(P)
Competencias C.	1.347	300,00	306,27	6,27*	0,13(P)

*Nota: * casos en que la diferencia de los promedios es significativa al 5%*

Tamaño del efecto: D – Despreciable, P-Pequeño, M-Mediano.

3.5.Diferencias de promedios en grupos de interés

Dentro de cada grado, puede que el efecto del formato sea más marcado en algunos grupos de estudiantes que en otros, ya que esto depende, en gran medida, de la familiaridad que los estudiantes tengan con el uso de computadores. La **Tabla 7** reporta el TE de la diferencia entre puntajes promedio por formato diferenciando por sexo, zona donde se ubica la sede educativa, y el sector educativo de la sede. Valores positivos del TE indican que el puntaje promedio es mayor para los estudiantes en PPL que en PEC, y valores negativos corresponden a un mayor puntaje para PEC.

Como se puede observar, en casi todos los casos se encuentra que la diferencia promedio está a favor de PPL, ya que se tiene un signo positivo para TE, excepto por unos pocos casos, en los cuales el TE es negativo pero insignificante (valor absoluto de TE menor a 0,1). Al comparar el TE para los dos sexos, se encuentran valores muy similares, indicando que el efecto de formato es muy similar para hombres y para mujeres, excepto por Competencias ciudadanas, en donde se encuentra un efecto más fuerte para las mujeres.

Al analizar el impacto del formato desagregando por zona, se encuentra que el efecto es simular en sedes urbanas y rurales para grado 3°. En grado 5° se encuentra que el TE incrementa alrededor de 0,08 puntos en sedes rurales para Lectura y Matemáticas, indicando un mayor efecto de formato en este tipo de escuelas. Se observa que el impacto del formato cambia de manera más drástica entre urbano y rural para grado 9°, especialmente en Lectura y Competencias ciudadanas. Esto indica que, al aplicar la prueba en los dos formatos, en general los estudiantes obtienen mayores puntajes en PPL (TE con valores positivos), pero ese efecto se ve más marcado en estudiantes de zonas rurales (valores más altos de TE), los cuales tienen una mayor reducción en sus puntajes al aplicar las pruebas electrónicas en comparación con las zonas urbanas.

Estudiantes que tienen menos familiaridad con el uso del computador tienen en general un mayor efecto por el formato de aplicación (Bennett et al., 2008), lo cual explica que estudiantes en zonas rurales tiendan a tener una mayor diferencia entre PEC y PPL, comparado con los estudiantes en zonas urbanas.

Con respecto al análisis desagregando por sector educativo, se encuentra que el efecto de formato es un poco más fuerte para el sector oficial en Lectura para grado 3°, así como para Matemáticas en grado 5°. Por el contrario, se encuentra que el impacto del formato es mayor en el sector privado para los estudiantes de 9° en la prueba de Competencias ciudadanas. En resumen, se concluye que los estudiantes obtienen mayores puntajes en PPL en general y que este impacto es más marcado en las zonas rurales y en colegios oficiales.

Tabla 7. Tamaño del efecto para la diferencia entre puntajes promedio según formato desagregando por sexo, zona y sector. Valores positivos del TE indican puntajes mayores para PPL y valores negativos corresponden a puntajes más altos para PEC.

Prueba	Sexo		Zona		Sector	
	Mujer	Hombre	Rural	Urbano	Oficial	Privado
Tercero						
Lectura	0,248	0,213	0,249	0,226	0,262	0,190
Matemáticas	0,069	0,032	-0,016	0,060	0,076	-0,013
Quinto						
Lectura	0,204	0,180	0,255	0,183	0,203	0,185
Matemáticas	0,087	0,096	0,171	0,076	0,117	0,021
Ciencias N.	-0,023	0,055	0,005	0,020	0,029	-0,010
Competencias C.	0,015	-0,073	0,032	-0,039	-0,017	-0,057
Noveno						
Lectura	0,301	0,302	0,416	0,294	0,302	0,301
Matemáticas	0,201	0,232	0,272	0,212	0,224	0,196
Ciencias N.	0,174	0,149	0,206	0,165	0,156	0,186
Competencias C.	0,165	0,085	0,223	0,128	0,077	0,270

*Nota: *Tamaño del Efecto: Despreciable (valores menores a 0,1), Pequeños (entre 0,1 y 0,3), Mediana (entre 0,3 y 0,6) o Grande (mayores a 0,6).*

3.6.Correlaciones en grupos de interés

De manera similar al análisis del TE desagregando por sexo, zona y sector educativo de la sede, se calcularon las correlaciones entre los puntajes en PEC y PPL en cada uno de estos grupos de estudiantes. Como se puede observar en la **Tabla 8**, la correlación aumenta con el grado de escolaridad en todos los grupos, lo cual indica que los estudiantes son más consistentes en cuanto a sus puntajes al aplicarles la prueba en los dos formatos en los grados más altos, como se mencionó anteriormente.

Por otro lado, las correlaciones son bastante similares entre los grupos de interés en cada grado-prueba, excepto por Competencias ciudadanas en 5° donde se encuentra una correlación más alta en las mujeres que en los hombres, y para Matemáticas en 5° también, que reporta una correlación mayor en zonas rurales que en las urbanas. La primera diferencia es particularmente amplia, por lo cual es importante revisar más en detalle por qué puede estar sucediendo esto en particular para el caso de Competencias ciudadanas en 5°.

Para comprender más en detalle los resultados, se puede revisar el texto de cada una de las preguntas para generar hipótesis que expliquen lo observado y también con ejercicios de pensar en voz alta, en los cuales los estudiantes resuelven los ítems y van detallando verbalmente el razonamiento que siguen para llegar a la respuesta.

Tabla 8. Correlaciones entre puntajes obtenidos por los estudiantes en PEC y PPL desagregando por sexo, zona y sector.

Prueba	Sexo		Zona		Sector	
	Mujer	Hombre	Rural	Urbano	Oficial	Privado
Tercero						
Lectura	0,58	0,59	0,63	0,58	0,55	0,57
Matemáticas	0,54	0,58	0,59	0,56	0,52	0,58
Quinto						
Lectura	0,66	0,62	0,67	0,63	0,63	0,59
Matemáticas	0,65	0,61	0,70	0,61	0,62	0,60
Ciencias N.	0,61	0,65	0,66	0,63	0,62	0,59
Competencias C.	0,61	0,48	0,58	0,54	0,52	0,58
Noveno						
Lectura	0,70	0,72	0,73	0,70	0,71	0,69
Matemáticas	0,71	0,76	0,77	0,73	0,72	0,77
Ciencias N.	0,68	0,70	0,71	0,69	0,68	0,70
Competencias C.	0,70	0,66	0,67	0,68	0,68	0,71

Pruebas Saber 3° 5° 7° 9°

04.

Conclusiones

Para el presente estudio se tuvo la misma población de estudiantes que presentaron la prueba en PEC y PPL, lo cual representó una gran ventaja, ya que cualquier diferencia que se observara en los resultados se puede atribuir al formato y no a posibles diferencias entre las muestras en los dos vehículos de aplicación. Se tuvieron alrededor de 5.000 estudiantes por grado que presentaron la prueba en los dos formatos.

La confiabilidad marginal fue alta para los dos formatos, lo cual indica que los evaluados se están midiendo de manera adecuada bajo los dos vehículos de aplicación. Además, la confiabilidad fue muy similar entre PEC y PPL, así que los dos formatos se comportan similarmente en cuanto al error de medición.

Respecto a los puntajes de los estudiantes en los dos vehículos de aplicación, se encontró que las correlaciones tienden a ser más altas cuando aumenta el grado de escolaridad, con valores de entre 0,56 y 0,59 para grado 3°, entre 0,56 y 0,64 para grado 5°, y entre 0,69 y 0,74 para grado 9°. Esto indica que, especialmente para 9°, estudiantes que obtuvieron un puntaje alto en la prueba en formato PEC en 2021 tienden a tener un puntaje alto en formato PPL en 2022. Lo mismo sucede para grados 3° y 5° pero la correlación es menos fuerte.

Al comparar los puntajes promedio de los estudiantes en los dos formatos, para 3° y 5° se encontraron diferencias pequeñas para Lectura con puntajes más altos en PPL que en PEC y diferencias despreciables en las demás pruebas. Esto se puede explicar teniendo en cuenta que las preguntas de la prueba de Lectura tienden a ser más largas, y el hecho de tener que desplazarse verticalmente a través de la pantalla para responder la pregunta puede llevar a menores puntajes en PEC. Similarmente, en el grado 9°, que es donde se tienen textos más largos en las preguntas, se encontraron mayores diferencias entre los promedios de PEC y PPL para las cuatro pruebas.

También se analizaron las diferencias entre el puntaje promedio en PEC y PPL para distintas poblaciones de interés, con el fin de determinar si la brecha es mayor en algunos grupos específicos de estudiantes. De esta manera, se concluyó que el efecto de formato es mayor en zonas rurales que urbanas, lo cual se puede explicar por la falta de familiaridad que tienen algunos estudiantes con el uso de computadores en las zonas rurales[3]. Como consecuencia, al aplicar la prueba en PEC, se puede estar castigando en mayor manera a los estudiantes de zonas rurales, y, por lo tanto, las brechas observadas en aplicaciones anteriores de Saber 3°, 5°, 9° en PPL entre rural y urbano se pueden ver acentuadas al realizar la prueba en PEC. Similarmente, se observa un mayor efecto de formato en colegios oficiales que en privados.

En síntesis, se encuentran algunas diferencias en cuanto a los puntajes promedio que obtienen los estudiantes dependiendo del formato. Según el tamaño del efecto, las diferencias no son despreciables en la mayoría de los casos, así que no se podría asegurar directamente que haya comparabilidad si se aplicara la prueba en los dos formatos simultáneamente. Además, las brechas se podrían ver acentuadas en los puntajes al comparar escuelas urbanas y rurales si se utiliza el formato PEC para evaluar a los estudiantes del país, así como al comparar instituciones oficiales y privadas.

Es importante llevar a cabo análisis cualitativos de los ítems para revisar en detalle el texto y las características de las preguntas con el fin de explicar más en detalle las razones por las cuales hay diferencias entre los resultados de los estudiantes al ser evaluados en PEC y en PPL. También se pueden desarrollar ejercicios de pensar en voz alta con los estudiantes para comprender mejor por qué se presentan dichas diferencias. Por último, se resalta que la aplicación de pruebas electrónicas puede presentar algunas dificultades en sedes alejadas de los centros urbanos, donde en algunos casos no se cuenta con servicios de electricidad que son un requerimiento mínimo para el funcionamiento de los computadores, por lo cual la transición hacia PEC no es solo un reto metodológico para la calificación sino que también logístico.

3 En la evaluación hay que tener en cuenta que las poblaciones son distintas, por lo cual no debería ser obligatorio tener las mismas condiciones de aplicación para todos los evaluados. El tener en cuenta sus diferencias y permitir que la aplicación se ajuste a sus características puede incrementar la validez de las pruebas (Sireci, 2020)

Pruebas Saber 3° 5° 7° 9°

05.

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.

Berman, A. I., Haertel, E. H., & Pellegrino, J. W. (2020). Comparability of Large-Scale Educational Assessments: Issues and Recommendations. Washington, DC: National Academy of Education

Bock, D. & Zimowsky (1997). Multiple group IRT. En Linden, W. & Hambleton, R. (Ed.), Handbook of modern item response theory. (pp. 433-448). Springer.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance. 28.

Carlbring, P., Brunt, S., Bohman, S., Austin, D., Richards, J., Öst, L.-G., & Andersson, G. (2007). Internet vs. Paper and pencil administration of questionnaires commonly used in panic/agoraphobia research. Computers in Human Behavior, 23(3), 1421-1434.

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. (1988). Statistical power analysis for the behavioral sciences (2nd ed). L. Erlbaum Associates.

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

Crooks, T. (Marzo, 2004) Tensions between assessment for learning and assessment for qualifications. Trabajo presentado en III Conference of the Association of Commonwealth Examinations and Accreditation Bodies, Nadi, Fiji.

Glass, V. & Hopkins, D. (1995). Statistical Methods in Education & Psychology. (3.a. ed). Alyn&Bacon.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21(4), 347–360. https://doi.org/10.1111/j.1745-3984.1984.tb01039.x

Guilford, J. P. (1953). The Correlation of an Item with a Composite of the Remaining Items in a Test. In Educational and Psychological Measurement (Vol. 13, Issue 1, pp. 87–93). SAGE Publications. https://doi.org/10.1177/001316445301300109

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

Heissel, J., Adam, E., Doleac, J., Figlio, D., & Meer, J. (2018). Testing, Stress, and Performance: How Students Respond Physiologically to High-Stakes Testing. Education Finance and Policy 2021; 16 (2): 183–208. doi: https://doi.org/10.1162/edfp_a_00306

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

Instituto Colombiano para la Evaluación de la Educación (s.f.). Evaluar para Avanzar. Preguntas frecuentes. [Infografía]. Icfes. Bogotá, Colombia: autor. Recuperado de https://www2.icfes.gov.co/documents/39286/2163563/Infografia+evaluar+para+avanzar+copia.pdf

Instituto Colombiano para la Evaluación de la Educación (19 de febrero de 2016). Resolución 126/2016. Por la cual se establece la escala de los resultados del Examen de Estado de Calidad de la Educación Superior Saber TyT y se dictan otras disposiciones. https://normograma.icfes.gov.co/docs/resolucion_icfes_0126_2016.htm

Instituto Colombiano para la Evaluación de la Educación (2019). 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 (2020). 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 (Icfes). (2022a). Informe nacional de resultados de las pruebas Saber 3°, 5° y 9°. Aplicación 2021. Bogotá, Colombia: autor.

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

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? Assessment in Education: Principles, Policy & Practice, 23(4), 495-518.

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

Kuder, G. F. y Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2(3), 151–160.

McCoy, S., Marks, P. V., Carr, C. L., & Mbarika, V. (2004). Electronic versus paper surveys: Analysis of potential psychometric biases. 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the, 8 pp.

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.

Organización para la Cooperación y el Desarrollo Económico (2013). Synergies for Better Learning. An International Perspective on Evaluation and Assessment. http://dx.doi.org/10.1787/9789264190658-en

Piaw, C. (2011). Comparisons Between Computer-Based Testing and Paper-Pencil Testing: Testing Effect, Test Scores. Testing Time and Testing Motivation. In Proceedings of the Informatics Conference at: University of Malaya (pp. 1-9).

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A Comparative Evaluation of Score Results from Computerized and Paper & Pencil Mathematics Testing in a Large Scale State Assessment Program. The Journal of Technology, Learning and Assessment, 3(6), Article 6.

Pommerich, M. (2004). Developing Computerized Versions of Paper-and-Pencil Tests: Mode Effects for Passage-Based Tests. The Journal of Technology, Learning and Assessment, 2(6), Article 6.

Programme for International Student Assessment (2016). PISA-Based test for schools, Technical report 2016. Vol. 1. The European Commission.

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. (Documentos de Trabajo Saber Investigar No. 1). Instituto Colombiano para la Evaluación de la Educación (Icfes). https://www.icfes.gov.co/web/guest/saber-investigar

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

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. Educational Measurement: Issues and Practice, 39(3), 100-105.

Way, W. D., Lin, C. H., & Kong, J. (2008). Maintaining Score Equivalence as Tests Transition Online: Issues, Approaches and Trends. 33.

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

Ziecky, M. (2014). An introduction to the use of evidence-centered design in test development. Psicología educativa, 20 (1), 79-87.

Sede Principal Dirección: Calle 26 No.69-76,Torre 2, Piso 17, Edificio Elemento, Bogotá - Cundinamarca

Código Postal: 111071 Línea de Atención: (601) 918 9022