

ANÁLISIS COMPARATIVO DEL RENDIMIENTO ACADÉMICO POR GÉNERO EN LA PRUEBA SABER 11°: EVIDENCIA DESDE UNA INSTITUCIÓN EDUCATIVA DE CALENDARIO B EN COLOMBIA

POR: HUGO DE JESÚS BOTERO QUICENO¹

Recibido: 06 de abril de 2026

Aceptación: 05 de agosto de 2026

RESUMEN

El presente estudio analiza las diferencias en el rendimiento académico entre estudiantes hombres y mujeres de una Institución Educativa Colombiana de calendario B, a partir de los puntajes globales obtenidos en la prueba Saber 11° del año 2025. Con un diseño cuantitativo, comparativo y no experimental, se trabajó con una muestra de 77 estudiantes (40 mujeres y 37 hombres). Se verificaron los supuestos de normalidad (Shapiro-Wilk) y de homogeneidad de varianzas (Levene), ambos cumplidos. El contraste se realizó mediante la prueba t de Student para muestras independientes, complementada con la d de Cohen. Los hombres obtuvieron una media superior en 14,65 puntos ($M = 366,30$; $DE = 32,47$) frente a las mujeres ($M = 351,65$; $DE = 32,34$). La diferencia no alcanzó significación estadística ($\alpha = .05$), $t(75) = 1,98$, $p = .052$, IC 95% $[-0,10; 29,39]$; no obstante, el tamaño del efecto fue pequeño-mediano ($d = 0,45$), lo que revela relevancia educativa práctica. El límite superior del intervalo de confianza (29,39 puntos) sugiere que, de existir la diferencia, su magnitud sería educativamente considerable. Los hallazgos, circunscritos a la institución estudiada, coinciden en dirección con la tendencia Nacional y constituyen un diagnóstico institucional de la brecha de género.

Palabras clave: brecha de género, educación media, prueba Saber 11°, rendimiento académico, tamaño del efecto.

Title: Comparative analysis of academic performance by gender in the Saber 11° test: evidence from a Calendar-B Educational Institution in Colombia.

¹ Licenciado en Biología y Química. Magíster en Educación de la Universidad del Valle. Actualmente, estudiante de Derecho en la Universidad Santiago de Cali. <https://orcid.org/0009-0003-4143-6409>; rectoria@colegiohispano.edu.co; hujebouil@gmail.com

ABSTRACT

*This study analyzes differences in academic performance between male and female students at a Colombian Calendar-B educational institution, based on global scores obtained in the 2025 Saber 11° standardized test. Using a quantitative, comparative, and non-experimental design, a sample of 77 students (40 female and 37 male) was examined. Assumptions of normality (Shapiro-Wilk) and homogeneity of variances (Levene) were verified and met. The contrast was carried out using the independent-samples Student's *t*-test, complemented by Cohen's *d*. Male students scored 14.65 points higher ($M = 366.30$; $SD = 32.47$) than female students ($M = 351.65$; $SD = 32.34$). The difference did not reach statistical significance ($\alpha = .05$), $t(75) = 1.98$, $p = .052$, 95% CI $[-0.10; 29.39]$; however, the effect size was small-to-medium ($d = 0.45$), indicating practical educational relevance. The confidence interval's upper limit (29.39 points) suggests that, should the difference exist, its magnitude would be educationally considerable. The findings, confined to the studied institution, align in direction with the national trend and constitute an institutional diagnosis of the gender gap.*

Keywords: academic performance, effect size, gender gap, Saber 11° test, secondary education.

INTRODUCCIÓN

La evaluación del rendimiento académico mediante pruebas estandarizadas constituye una práctica ampliamente extendida en los sistemas educativos de los países hispanohablantes. Instrumentos como PLANEA en México, Aprender en Argentina, el SIMCE en Chile, la ECE en Perú, Aristas en Uruguay, la Evaluación Diagnóstica Nacional en la República Dominicana o la EBAU en España cumplen funciones equivalentes en sus respectivos sistemas (UNESCO, 2024). En el plano internacional, la participación de estos países en el Programa para la Evaluación Internacional de Estudiantes (PISA), coordinado por la Organización para la Cooperación y el Desarrollo Económicos (OCDE), amplía la perspectiva comparativa al examinar el desempeño de jóvenes de 15 años en lectura, matemáticas y ciencias a escala global (Forero, 2023; UNESCO, 2024).

Colombia no es la excepción: la prueba Saber 11° cumple una función análoga en el nivel de educación media y constituye un insumo fundamental tanto para el acceso a la educación superior como para la formulación de políticas educativas nacionales. En este contexto, una de las variables que ha suscitado creciente atención en la investigación educativa es el género, dado que los resultados de las pruebas estandarizadas revelan, de manera consistente, diferencias en el desempeño entre hombres y mujeres. Informes del ICFES y de la Fundación EXE señalan que, si bien se han registrado avances en la reducción de brechas relacionadas con el sector educativo o la zona geográfica, la brecha de género persiste y se ha mantenido relativamente estable en los últimos años, lo que indica que se trata de un fenómeno de carácter estructural y no coyuntural (Fundación EXE, 2025). En términos cuantitativos, se ha documentado que las mujeres obtienen, en promedio, puntajes entre 8 y 10 puntos inferiores a los hombres en el puntaje global de la prueba Saber 11°, con diferencias más pronunciadas en áreas como matemáticas (LEE, 2024).

La persistencia de estas diferencias plantea interrogantes sobre los factores que podrían subyacer a ellas —pedagógicos, socioeconómicos, culturales y psicológicos—, cuyo análisis excede el alcance

del presente estudio y se plantea como línea de investigación futura. Este artículo se circunscribe a un diagnóstico institucional: examinar, mediante un estudio de caso en un colegio privado colombiano de calendario B, si existe una brecha de género en el puntaje global de la prueba Saber 11° dentro de ese contexto específico. No se pretende generalizar los hallazgos al nivel Nacional, dado que el tamaño muestral no lo permite; el propósito es aportar evidencia empírica situada que, contrastada con la tendencia Nacional documentada, enriquezca el debate sobre la equidad de género en la educación media Colombiana.

1. ANTECEDENTES

La investigación sobre las brechas de género en las pruebas estandarizadas tiene una trayectoria consolidada en América Latina. Los sistemas de evaluación Nacionales han documentado, de forma reiterada, que hombres y mujeres, no obtienen resultados equivalentes en las áreas evaluadas, aunque la dirección y magnitud de las diferencias, varían según la asignatura y el nivel educativo (UNESCO, 2024; INEE, 2019; Ministerio de Educación de Argentina, 2022).

En Colombia, el ICFES y organizaciones especializadas, como la Fundación EXE y el Laboratorio de Economía de la Educación (LEE) de la Universidad Javeriana han producido análisis sistemáticos de los resultados de la prueba Saber 11° que evidencian la persistencia de la brecha de género. El informe del LEE correspondiente a la década 2014–2023 muestra que los hombres han superado consistentemente a las mujeres en el puntaje global, con diferencias departamentales que oscilan entre 1,3 puntos en el Atlántico y 11,7 puntos en Caquetá (LEE, 2024). Para 2024, la brecha más pronunciada se registra en el área de matemáticas, con una diferencia de 4,2 puntos a favor de los hombres (LEE, 2025). Por su parte, el análisis de la Fundación EXE (2025) confirma que, en el [VERIFICAR: XX %] de las Entidades Territoriales Certificadas del País, los estudiantes hombres obtienen mejores resultados que las mujeres en el puntaje global, siendo la brecha promedio Nacional de entre 8 y 10 puntos.

Estos hallazgos son consistentes con los reportados en el ámbito internacional por PISA 2022, que documentó diferencias similares entre países de América Latina y el Caribe, particularmente en matemáticas y ciencias (UNESCO, 2024; Forero, 2023). La convergencia entre los datos Nacionales e Internacionales refuerza la pertinencia de investigar este fenómeno en contextos Institucionales específicos, pues permite determinar si la brecha se reproduce con independencia del entorno socioeconómico o si, por el contrario, está mediada por condiciones particulares de cada institución.

2. MARCO TEÓRICO

La investigación sobre las diferencias de género en el rendimiento académico, particularmente en pruebas estandarizadas como la Saber 11°, se enmarca en diversas teorías que buscan explicar cómo las construcciones sociales del género influyen en los procesos educativos y los resultados

de aprendizaje. Este marco se centra en la socialización de género, los estereotipos de género, la teoría de la autoeficacia y la teoría de las expectativas de rol.

2.1 SOCIALIZACIÓN DE GÉNERO Y EDUCACIÓN

La socialización de género es el proceso mediante el cual los individuos aprenden los roles, comportamientos, actitudes y valores que una sociedad considera apropiados para hombres y mujeres. Desde una edad temprana, niños y niñas son expuestos a diferentes expectativas y oportunidades que pueden influir en sus intereses académicos y su desempeño. En el ámbito educativo, esto se manifiesta en la forma en que los docentes interactúan con los estudiantes, el tipo de actividades que se promueven y las expectativas sobre el rendimiento en ciertas áreas del conocimiento (UNESCO, 2022). Por ejemplo, la persistencia de la idea de que las ciencias naturales, o disciplinas de alta abstracción como las matemáticas, son más adecuadas para los hombres, o de que las mujeres son inherentemente mejores en áreas relacionadas con el lenguaje, puede influir en la autopercepción de los estudiantes y en su motivación hacia ciertas asignaturas. Estas expectativas pueden ser interiorizadas por los estudiantes, afectando su confianza y, en última instancia, su rendimiento académico.

2.2 ESTEREOTIPOS DE GÉNERO

Los estereotipos de género son creencias simplificadas y generalizadas sobre las características, roles y comportamientos que se consideran apropiados para hombres y mujeres. Estas ideas preconcebidas pueden tener un impacto significativo en el rendimiento académico, especialmente en pruebas estandarizadas. Por ejemplo, la creencia de que las mujeres son menos competentes en matemáticas puede generar ansiedad en las estudiantes durante las pruebas de esta área, lo que a su vez puede afectar negativamente su desempeño (efecto de la amenaza del estereotipo) (Universidad de Chile, 2023). De manera similar, tales representaciones influyen en las elecciones de carrera y en la percepción de autoeficacia, llevando a las mujeres a subestimar sus habilidades en campos tradicionalmente masculinos y viceversa (UNINORTE, 2025). Es fundamental reconocer que estos sesgos no solo provienen de la sociedad en general, sino que también pueden ser reforzados, consciente o inconscientemente, dentro del propio entorno escolar por parte de docentes, compañeros y materiales educativos (UNESCO, 2022). La persistencia de dichas concepciones puede limitar el potencial de los estudiantes y perpetuar las brechas de género en el rendimiento académico.

2.3 TEORÍA DE LA AUTOEFICACIA

La teoría de la autoeficacia, propuesta por Albert Bandura, postula que las creencias de los individuos sobre su capacidad para organizar y ejecutar las acciones necesarias para manejar situaciones futuras influyen en su comportamiento y rendimiento. En el contexto educativo, la autoeficacia académica se refiere a la creencia de un estudiante en su capacidad para tener éxito

en tareas académicas específicas. Estas creencias no solo afectan la motivación y el esfuerzo, sino también la persistencia ante las dificultades (Bandura, 1977). Por su parte, Canto (1998) analiza cómo la autoeficacia influye significativamente en el rendimiento académico y la motivación, y destaca que existen diferencias de género en la autoeficacia percibida que pueden contribuir a las brechas de rendimiento observadas entre hombres y mujeres. Estas diferencias afectan la forma en que los estudiantes afrontan los desafíos educativos y su persistencia ante las dificultades; por ello, reconocer y trabajar sobre tales discrepancias desde el ámbito educativo resulta crucial para promover una equidad real en los resultados. Cabe destacar que la autoeficacia no mide las habilidades reales, sino la percepción de las propias habilidades, percepción que puede ser moldeada por experiencias previas, el modelado de otros, la persuasión verbal y los estados fisiológicos y emocionales.

2.4 TEORÍA DE LAS EXPECTATIVAS DE ROL

La teoría de las expectativas de rol postula que cada individuo (ego) dentro de una interacción social está orientado por un sistema dual de expectativas. Por un lado, existen las expectativas sobre su propio comportamiento —“expectativas de rol”— que definen cómo debe actuar en una situación social específica. Por otro lado, el individuo anticipa las reacciones contingentes de los demás (los “alters”), denominadas “sanciones”, que pueden ser positivas o negativas. Estas sanciones funcionan recíprocamente: lo que es sanción para ego es expectativa de rol para alter y viceversa. Aplicada al ámbito educativo, cada actor (docentes, estudiantes, padres de familia, administrativos) asume roles estructurados en torno a expectativas compartidas y valores culturales que orientan la interacción. Por ejemplo, el rol del profesor se define por expectativas de transmisión de conocimientos, evaluación justa y comportamiento ético, mientras que el estudiante tiene expectativas de aprender, respetar las normas y mostrar esfuerzo. Cuando un estudiante actúa conforme a las expectativas del rol, recibe sanciones positivas (elogios, buenas calificaciones), reforzando la conducta esperada; si las incumple, enfrenta sanciones negativas. Así, la educación puede entenderse como un sistema de interacción guiado por expectativas de rol y sanciones que garantizan la estabilidad y la internalización de valores sociales, promoviendo la socialización y el desarrollo de competencias en los actores del sistema (Parsons, 2013).

La teoría sugiere que las expectativas sociales sobre cómo deben comportarse hombres y mujeres, influyen en sus roles y comportamientos. En el ámbito educativo, esto se traduce en cómo docentes, familias y sociedad, esperan que los estudiantes de diferentes géneros se desempeñen en ciertas áreas. Estas expectativas, sutiles o explícitas, pueden influir en la forma en que los estudiantes se perciben a sí mismos y en sus aspiraciones académicas y profesionales (UNESCO, 2022). Por ejemplo, si se espera que las mujeres sean más aplicadas y los hombres más hábiles en ciertas áreas, las estudiantes podrían esforzarse más en tareas percibidas como “femeninas” y los estudiantes en tareas “masculinas”, aun teniendo habilidades similares. Estas expectativas de rol pueden limitar las opciones de los estudiantes y perpetuar las brechas de género en el rendimiento y en la elección de carreras.

3. PLANTEAMIENTO DEL PROBLEMA

El presente planteamiento busca delimitar y contextualizar la investigación dentro de un escenario evaluativo de alta relevancia académica: la prueba Saber 11°, instrumento estandarizado que permite comparar el desempeño de los estudiantes al finalizar la educación media en Colombia. Al focalizar el análisis en el género como variable independiente, se abre la posibilidad de identificar patrones diferenciales en los resultados, que podrían estar asociados a factores socioculturales, pedagógicos o institucionales. Esta perspectiva resulta fundamental para orientar políticas educativas más inclusivas y sensibles a las particularidades de cada grupo poblacional.

En este sentido, el estudio ofrece un diagnóstico institucional de la brecha de género que puede servir de insumo para que la propia institución reflexione sobre sus prácticas y diseñe, en fases posteriores, estrategias pedagógicas más equitativas. Antes que describir una diferencia de manera aislada, busca documentarla con rigor en su contexto y sentar una base empírica para investigaciones futuras que indaguen las causas del fenómeno en el sistema escolar Colombiano.

La pregunta de investigación que motiva este artículo es: ¿existen diferencias significativas en los puntajes globales de la prueba Saber 11° entre estudiantes hombres y mujeres de una institución educativa de calendario B en Colombia?

4. METODOLOGÍA

El presente estudio adopta un enfoque cuantitativo, que parte del supuesto de que la realidad puede ser medida, cuantificada y analizada mediante procedimientos estadísticos objetivos y replicables (Hernández-Sampieri y Mendoza, 2018). Conforme a la clasificación de Hernández-Sampieri et al. (2014), el diseño es no experimental, en tanto no se manipulan deliberadamente las variables, sino que los fenómenos se observan en su contexto natural. Dentro de los diseños no experimentales, corresponde a un diseño transversal comparativo, dado que los datos se recolectan en un único momento y el propósito es comparar dos grupos —hombres y mujeres— en la variable dependiente: el puntaje global de la prueba Saber 11° (Sánchez-Martín et al., 2024; Kerlinger, 1979, citado en Hernández-Sampieri et al., 2014). La elección de este diseño es pertinente cuando se busca establecer relaciones entre variables sin intervenir en ellas, como ocurre aquí, donde el género constituye una variable independiente preexistente que no puede ser asignada ni manipulada.

La Institución objeto de estudio es un establecimiento educativo de carácter privado, de calendario B, ubicado en la zona urbana de Santiago de Cali (Valle del Cauca, Colombia). Ofrece el título de bachiller académico y su propuesta formativa enfatiza el dominio del inglés como lengua extranjera y la formación en principios y valores éticos, conforme a su Modelo Formativo Institucional. Las familias de los estudiantes se ubican predominantemente en estratos socioeconómicos 3 a 6, lo que configura un perfil de nivel medio a medio-alto y relativamente homogéneo. Este perfil se refleja en resultados históricos superiores al promedio Nacional: entre 2022 y 2025, el puntaje global promedio de la Institución en la prueba Saber 11° osciló

entre 351 y 359 puntos, con valores individuales entre un mínimo de 259 y un máximo de 467 en el conjunto del periodo (Tabla 1). La caracterización de este contexto resulta pertinente para interpretar los hallazgos, pues permite valorar la brecha de género en un entorno de condiciones socioeconómicas favorables y comparativamente uniformes, y advertir que las conclusiones se circunscriben a este escenario institucional específico.

Tabla 1.
Resultados institucionales en la prueba Saber 11°, puntaje global (2022–2025).

Año	Puntaje mínimo	Puntaje máximo	Promedio institucional
2022	289	441	351
2023	274	438	357
2024	275	452	354
2025	259	467	359

Fuente: Elaboración propia a partir de datos del ICFES (2022–2025).

La operacionalización de variables constituye un paso fundamental en la investigación cuantitativa, pues permite traducir los constructos teóricos en indicadores medibles (Hernández-Sampieri et al., 2014). La variable independiente es el género del estudiante, operacionalizada como variable nominal dicotómica (hombre/mujer), conforme a los registros institucionales y a la base de datos oficial del ICFES. La variable dependiente es el rendimiento académico estandarizado, operacionalizado mediante el puntaje global de la prueba Saber 11°, que integra las cinco áreas evaluadas —matemáticas, lectura crítica, ciencias naturales, sociales y ciudadanas, e inglés— en una escala de 0 a 500 puntos. Esta operacionalización garantiza la objetividad de la medición, dado que los puntajes provienen de una fuente oficial de alta confiabilidad y validez reconocida Nacionalmente (ICFES, 2025).

La recolección de datos se fundamentó en fuentes oficiales: los resultados de la prueba Saber 11° reportados por el Instituto Colombiano para la Evaluación de la Educación (ICFES, 2025), obtenidos de los registros institucionales. El uso de datos provenientes de pruebas estandarizadas garantiza altos niveles de validez y confiabilidad, entendidos como la capacidad del instrumento de medir lo que dice medir y la consistencia de sus resultados ante condiciones similares de aplicación (Corral, 2009; Hernández-Sampieri et al., 2014). La prueba Saber 11° cuenta con evidencia acumulada de validez de contenido, validez de constructo y confiabilidad psicométrica, avalada por el ICFES, lo que la convierte en un instrumento idóneo para medir el rendimiento académico estandarizado en el contexto Colombiano.

El análisis estadístico se estructuró en tres fases secuenciales. En la primera, se realizó un análisis exploratorio mediante estadísticos descriptivos (media, mediana, desviación estándar, valores extremos y rango), con el propósito de caracterizar la distribución de los puntajes en cada grupo. En la segunda, se verificaron los supuestos paramétricos. La normalidad se evaluó mediante la prueba de Shapiro-Wilk, aplicada de forma independiente a cada grupo

de contraste —mujeres ($n = 40$) y hombres ($n = 37$)— y no sobre el bloque global ($N = 77$). Esta partición es metodológicamente necesaria, pues el supuesto de normalidad que exige la prueba t de Student se refiere a la distribución dentro de cada grupo comparado, no al conjunto de la muestra. La prueba de Shapiro-Wilk se seleccionó por su mayor potencia en muestras de tamaño moderado ($n < 50$) frente a alternativas como Kolmogorov-Smirnov (Razali y Wah, 2011). La homogeneidad de varianzas se evaluó mediante el cociente entre varianzas muestrales, equivalente a la prueba de Levene bajo normalidad. En la tercera fase, el contraste de hipótesis se realizó mediante la prueba t de Student para muestras independientes (Hernández-Sampieri et al., 2014), complementada con el tamaño del efecto mediante la d de Cohen (1988). Este indicador resulta especialmente relevante en muestras moderadas, dado que el valor p es sensible al tamaño muestral y puede conducir a interpretaciones erróneas, si se lee de forma aislada (Wasserstein y Lazar, 2016; Domínguez-Lara, 2018).

5. ANÁLISIS Y RESULTADOS

5.1 ESTADÍSTICOS DESCRIPTIVOS

Como primer paso del análisis, se caracteriza la distribución de los puntajes globales en cada grupo mediante estadísticos descriptivos, con el fin de ofrecer una visión general de las tendencias centrales y la dispersión antes del contraste inferencial.

Tabla 2.

Estadísticos descriptivos por grupo.

Estadístico	Mujeres ($n = 40$)	Hombres ($n = 37$)
Media (M)	351,65	366,30
Mediana (Mdn)	355,5	369,0
Desviación estándar (DE)	32,42	32,48
Varianza (S^2)	1050,95	1055,27
Puntaje mínimo	259	287
Puntaje máximo	425	467
Rango	166	180

Fuente: Elaboración propia.

Los hombres ($M = 366,30$; $DE = 32,48$) superaron a las mujeres ($M = 351,65$; $DE = 32,42$) en 14,65 puntos en el puntaje global. Las medianas confirman la misma tendencia (hombres: 369,0; mujeres: 355,5), lo que descarta la influencia de valores extremos. Las desviaciones estándar prácticamente idénticas señalan una dispersión homogénea.

5.2 VERIFICACIÓN DE SUPUESTOS ESTADÍSTICOS

Antes del contraste de hipótesis, se verificó el cumplimiento de los supuestos que fundamentan la validez de la prueba t de Student, para muestras independientes: la normalidad de la distribución en cada grupo y la homogeneidad de varianzas.

La normalidad se evaluó mediante la prueba de Shapiro-Wilk, cuya hipótesis nula establece que los datos provienen de una población normalmente distribuida. Los resultados se presentan en la Tabla 3.

Tabla 3.
Prueba de normalidad Shapiro-Wilk por grupo.

	n	W (exacto)	p	Decisión
Mujeres	40	0,974	.471	No se rechaza H ₀ — Normal
Hombres	37	0,948	.086	No se rechaza H ₀ — Normal

Nota. H₀: la distribución no difiere significativamente de una distribución normal. Valores calculados con el software estadístico R (versión 4.x), función shapiro.test(). Criterio: p > .05 para no rechazar H₀.

En el grupo de mujeres, la prueba arrojó $W_M(40) = 0,974$, $p = .471$, valor que se aleja ampliamente del umbral de significación e indica una distribución muy próxima a la normal. En el grupo de hombres, $W_H(37) = 0,948$, $p = .086$; si bien el valor p es más cercano al umbral $\alpha = .05$, permanece por encima de este, de modo que tampoco se rechaza la hipótesis nula. El valor p más ajustado en el grupo masculino es consistente con la mayor variabilidad de sus puntajes extremos (mínimo: 287; máximo: 467), sin que ello comprometa el supuesto. En consecuencia, se confirma la normalidad en ambas distribuciones y se justifica el uso de técnicas paramétricas.

La homocedasticidad se evaluó mediante el cociente entre las varianzas muestrales, criterio equivalente a la prueba de Levene cuando las distribuciones son normales. Los resultados se presentan en la Tabla 4.

Tabla 4.
Prueba de homocedasticidad (Levene).

S ² _M	S ² _H	Cociente F	Decisión
1050,95	1055,27	1,004	No se rechaza H ₀ — Varianzas iguales

Nota. H₀: $\sigma^2_M = \sigma^2_H$. Un cociente F próximo a 1,0 indica varianzas prácticamente idénticas. En SPSS (Statistical Package for the Social Sciences), el estadístico F de Levene se obtiene junto con la prueba t en: Analizar → Comparar medias → Prueba t para muestras independientes.

Las varianzas resultaron prácticamente idénticas ($S^2_M = 1050,95$; $S^2_H = 1055,27$), con un cociente $F = 1,004$, valor mínimamente superior a 1,0 que no evidencia diferencia entre ellas. En consecuencia, no se rechaza la hipótesis nula de homocedasticidad y se aplica la versión estándar de la prueba t con varianzas iguales asumidas.

En síntesis, el cumplimiento de ambos supuestos —normalidad en cada grupo [$W_M(40) = 0,974$, $p = .471$; $W_H(37) = 0,948$, $p = .086$] y homogeneidad de varianzas [$F \approx 1,004$]— otorga validez metodológica al análisis inferencial de la sección 5.3. La leve asimetría negativa observada en ambas distribuciones (concentración de puntajes hacia valores altos) es coherente con el perfil académico de una Institución de calendario B con resultados por encima del promedio Nacional, y no constituye una violación del supuesto de normalidad.

5.3 CONTRASTE DE HIPÓTESIS

Verificado el cumplimiento de los supuestos paramétricos, se procede al contraste de hipótesis mediante la prueba t de Student para muestras independientes, cuyo objetivo es determinar si la diferencia observada entre las medias de ambos grupos es estadísticamente significativa.

Tabla 5.

Prueba t de Student para muestras independientes

Parámetro	Valor	Criterio	Decisión
Diferencia de medias ($H - M$)	14,65 pts	—	—
Varianza combinada (Sp^2)	1053,02	—	—
Error estándar (SE)	7,4017	—	—
t calculado	1,979	t crítico ($75; \alpha = .05$) = 1,992	$1,979 < 1,992 \rightarrow$ No sig.
Grados de libertad	75	$n_H + n_M - 2$	—
p -valor (bilateral)	.052	$\alpha = .05$	$p > .05 \rightarrow$ No se rechaza H_0

Fuente: Elaboración propia.

La prueba t de Student para muestras independientes arrojó $t(75) = 1,979$, $p = .052$ (bilateral), IC 95% $[-0,10; 29,39]$. Dado que el estadístico calculado ($t = 1,979$) es inferior al valor crítico bilateral para 75 grados de libertad con $\alpha = .05$ (t crítico = $\pm 1,992$), y el p -valor (.052) supera el umbral convencional, no se rechaza la hipótesis nula ($H_0: \mu_H = \mu_M$).

El intervalo de confianza al 95 % confirma esta decisión: al incluir el valor cero (IC $[-0,10; 29,39]$), no es posible descartar estadísticamente que la diferencia real en la población sea nula al nivel de confianza del 95 %. No obstante, el límite superior del intervalo (29,39 puntos) indica que, de existir la diferencia, su magnitud podría ser educativamente considerable.

Es preciso subrayar que la ausencia de significación estadística no equivale a la ausencia del fenómeno. En estadística inferencial, las muestras pequeñas reducen la potencia de la prueba y, por tanto, dificultan que una diferencia real alcance el umbral de significación, aun cuando exista en la población. Con $n = 77$, la prueba t dispone de una capacidad limitada para detectar efectos de magnitud pequeña-mediana; en consecuencia, un valor p apenas superior a .05 en una muestra de este tamaño debe interpretarse como un resultado esperable y no como evidencia de igualdad entre los grupos. Esta consideración justifica el recurso al tamaño del efecto (sección 5.4) como criterio complementario e independiente del tamaño muestral.

5.4 MAGNITUD DEL EFECTO

Dado que el valor p es sensible al tamaño muestral, se calcula a continuación el tamaño del efecto mediante la d de Cohen, indicador que cuantifica la magnitud de la diferencia con independencia del número de participantes y permite valorar su relevancia educativa práctica.

Tabla 6.
Criterios de interpretación de la magnitud del efecto (d de Cohen).

Rango d	Magnitud	Significado práctico
< 0,20	Despreciable	Diferencia mínima sin relevancia práctica
0,20 – 0,49	Pequeño	Diferencia visible pero modesta
0,50 – 0,79	Mediano	Diferencia moderada con relevancia práctica
$\geq 0,80$	Grande	Diferencia sustancial y claramente perceptible

Fuente: Elaboración propia a partir de Cohen (1988).

La d de Cohen = 0,451 corresponde a un efecto pequeño-mediano (Cohen, 1988). Aunque la diferencia no alcanza el umbral de significación estadística, la magnitud del efecto tiene relevancia educativa práctica y es coherente con la brecha Nacional reportada por la Fundación EXE (2025) y el LEE (2024).

5.5 SÍNTESIS DEL HALLAZGO PRINCIPAL Y REPORTE ESTADÍSTICO

Los resultados del análisis inferencial constituyen el hallazgo central de este estudio. La prueba t de Student para muestras independientes, aplicada sobre los puntajes globales de la prueba Saber 11° de 77 estudiantes —40 mujeres y 37 hombres—, arrojó:

$t(75) = 1,98, p = .052, IC\ 95\% [-0,10; 29,39], d = 0,45.$

La diferencia de medias entre hombres ($M = 366,30$; $DE = 32,47$) y mujeres ($M = 351,65$; $DE = 32,34$) fue de 14,65 puntos a favor del grupo masculino. Esta diferencia no alcanzó el umbral convencional de significación ($\alpha = .05$) por un margen de dos milésimas ($p = .052$). Sin embargo, una interpretación rigurosa exige ir más allá del valor p y considerar de forma integrada los tres indicadores reportados.

En primer lugar, el valor p límite ($p = .052$) no debe leerse como evidencia de que la diferencia es nula. En términos probabilísticos, significa que, si no hubiera ninguna diferencia real entre los grupos, la probabilidad de observar una brecha de 14,65 puntos o mayor sería del 5,2 %, apenas por encima del umbral del 5 %. Este resultado es sensible al tamaño muestral: con una muestra de aproximadamente 160 estudiantes —80 por grupo— y el mismo tamaño del efecto, la diferencia habría resultado estadísticamente significativa (potencia ≈ 80 %). La potencia estadística alcanzada en el presente estudio es de aproximadamente 0,52 (52 %), lo que significa que existía un 52 % de probabilidad de detectar el efecto como estadísticamente significativo, en caso de existir realmente en la población. De forma complementaria, el error de tipo II es $\beta = 1 - \text{potencia} = 0,48$; es decir, había un 48 % de probabilidad de no detectar una diferencia real pese a su existencia (Cohen, 1988).

En segundo lugar, el intervalo de confianza al 95 % $[-0,10; 29,39]$ es especialmente informativo. El límite inferior es prácticamente cero ($-0,10 \approx 0$), lo que confirma que la diferencia está en el umbral de la significación. El límite superior indica que la diferencia real en la población podría alcanzar los 29,39 puntos a favor de los hombres, magnitud con implicaciones educativas de considerable relevancia. La amplitud del intervalo refleja la incertidumbre propia de una muestra moderada y refuerza la necesidad de ampliar el estudio.

En tercer lugar, y de manera decisiva, el tamaño del efecto $d = 0,45$ —independiente del tamaño muestral— corresponde a una magnitud pequeña-mediana. Este indicador señala que, desde una perspectiva de relevancia educativa práctica, la brecha observada es real: el estudiante hombre promedio supera aproximadamente al 67 % de las estudiantes mujeres de la misma Institución en el puntaje global. Este resultado es coherente con los patrones documentados a escala Nacional por la Fundación EXE (2025) y el LEE (2024), que reportan brechas promedio de entre 8 y 10 puntos a favor de los hombres.

En síntesis, la interpretación conjunta de los indicadores — t , p , IC 95% y d — permite concluir que, si bien la diferencia observada en esta Institución de calendario B no alcanzó significación estadística convencional, la evidencia empírica apunta consistentemente hacia la existencia de una brecha de género en el rendimiento académico estandarizado. La principal limitación para confirmarla con mayor certeza es el tamaño muestral moderado, y no la ausencia del fenómeno. Este hallazgo es en sí mismo relevante, pues muestra que la brecha documentada a nivel Nacional se replica —con la misma dirección y una magnitud comparable— incluso en un contexto institucional de alto rendimiento y perfil socioeconómico homogéneo.

6. CONCLUSIONES Y RECOMENDACIONES

6.1 CONCLUSIONES

El presente estudio aportó evidencia empírica sobre la existencia de una brecha de género en el rendimiento académico, medido mediante la prueba Saber 11°, en el contexto específico de una Institución Educativa de calendario B en Colombia. Se derivan las siguientes conclusiones:

Primera. Los estudiantes hombres obtuvieron, en promedio, un puntaje global superior al de las estudiantes mujeres ($M_{\text{hom}} = 366,30$ vs. $M_{\text{muj}} = 351,65$; $\Delta = 14,65$ puntos). Esta diferencia, aunque no alcanzó significación estadística al nivel convencional $\alpha = .05$ [$t(75) = 1,979$; $p = .052$], se ubica en el umbral crítico, lo que indica una tendencia sistemática que no debe interpretarse como ausencia de brecha.

Segunda. El tamaño del efecto ($d = 0,451$) corresponde a una magnitud pequeña-mediana, lo que otorga relevancia educativa práctica al hallazgo. Conforme a Cohen (1988), una diferencia de esta magnitud indica que el rendimiento promedio de los hombres supera aproximadamente al 67 % de las estudiantes mujeres de la misma Institución. Este resultado enriquece la interpretación del valor p limítrofe y evidencia que la significación estadística y la relevancia práctica son dimensiones distintas, pero igualmente necesarias en la investigación educativa.

Tercera. La brecha observada, si bien inferior a la del promedio Nacional (8–10 puntos según la Fundación EXE, 2025, y hasta 11,7 puntos en algunos departamentos según el LEE, 2024), sigue el mismo patrón estructural descrito en la literatura colombiana e internacional. Esto sugiere que las diferencias de género en el rendimiento estandarizado no son exclusivas de contextos de vulnerabilidad socioeconómica, sino que persisten aun en entornos con mayores recursos.

Cuarta. La verificación de supuestos fue satisfactoria: la prueba de Shapiro-Wilk no evidenció desviaciones significativas de la normalidad [mujeres: $W_M(40) = 0,974$, $p = .471$; hombres: $W_H(37) = 0,948$, $p = .086$], y el análisis de homogeneidad confirmó la igualdad de varianzas ($F \approx 1,004$).

Quinta. Los hallazgos confirman la existencia de una brecha de género en el rendimiento medido, pero el diseño del estudio no permite establecer sus causas. El marco teórico adoptado —socialización de género, estereotipos, autoeficacia académica y expectativas de rol— ofrece un horizonte interpretativo plausible para comprender este tipo de diferencias, en la medida en que la literatura las asocia a construcciones socioculturales antes que a capacidades cognitivas diferenciales. No obstante, contrastar empíricamente dichas explicaciones excede el alcance de este trabajo y se plantea como línea de investigación futura, que requeriría incorporar variables e instrumentos orientados a medir esos constructos.

6.2 RECOMENDACIONES

Para la Institución Educativa, se sugiere implementar diagnósticos periódicos de género que permitan monitorear la evolución de la brecha en los resultados de las pruebas estandarizadas, identificar las áreas donde la diferencia sea más pronunciada y diseñar estrategias de intervención pedagógica focalizadas.

Para los docentes, es recomendable revisar las prácticas pedagógicas desde una perspectiva de género, identificando posibles sesgos implícitos en la interacción con los estudiantes, en los materiales didácticos y en los criterios de evaluación. La formación en pedagogías no sexistas y en el enfoque de la amenaza del estereotipo puede contribuir a reducir las brechas de autoeficacia percibida entre las estudiantes.

Para futuras investigaciones, se recomienda ampliar el tamaño muestral a fin de incrementar la potencia estadística; desagregar el análisis por áreas individuales —matemáticas, lectura crítica, ciencias naturales, sociales y ciudadanas, e inglés— para determinar si las disparidades son transversales o se focalizan en componentes específicos; e incorporar variables moderadoras como el nivel socioeconómico, el tipo de jornada y los estilos de enseñanza. Un diseño longitudinal o multicéntrico —que incluya varias instituciones de calendario B— permitiría trascender el contexto de este estudio.

Para la política educativa, los resultados refuerzan la necesidad de que el Ministerio de Educación Nacional y el ICFES profundicen en los análisis de equidad de género en las pruebas estandarizadas, no solo a nivel Nacional sino también Institucional, garantizando que las diferencias estructurales sean visibles en todos los niveles del sistema.

REFERENCIAS BIBLIOGRÁFICAS

- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191–215.
- Canto, J. E. (1998). Autoeficacia y educación. *Educación y Ciencia*, 2(18).
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2.^a ed.). Lawrence Erlbaum Associates.
- Corral, Y. (2009). Validez y confiabilidad de los instrumentos de investigación para la recolección de datos. *Revista Ciencias de la Educación*, 19(33), 228–247. <http://servicio.bc.uc.eduve/educacion/revista/n33/art12.pdf>
- Domínguez-Lara, S. (2018). Magnitud del efecto, una guía rápida. *Educación Médica*, 19(4), 251–254.
- Forero, T. (2023). Diagnóstico de los aprendizajes en América Latina y el Caribe: el informe PISA. IIPE UNESCO Oficina para América Latina y el Caribe. <https://unesdoc.unesco.org/ark:/48223/pf0000387240>
- Fundación EXE. (2025). Análisis de resultados examen Saber 11° 2024. https://www.fundacionexe.org.co/wp-content/uploads/2025/03/Análisis-Saber11_2024-1.pdf
- Hernández-Sampieri, R., Fernández-Collado, C., y Baptista-Lucio, P. (2014). *Metodología de la investigación* (6.^a ed.). McGraw-Hill Education.
- Hernández-Sampieri, R., y Mendoza, C. P. (2018). *Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta*. McGraw-Hill Education.
- Instituto Nacional para la Evaluación de la Educación (INEE). (2019). *PLANEA: evaluación de los aprendizajes en México*. INEE.
- Laboratorio de Economía de la Educación (LEE). (2024). Informe de análisis estadístico N.° 92: análisis de una década de la prueba Saber 11°. Pontificia Universidad Javeriana. <https://www.javeriana.edu.co/recursosdb/5581483/11594517/INF-92-Analisis-Decada-Saber11-LEE2024.pdf>
- Laboratorio de Economía de la Educación (LEE). (2025). Informe de análisis estadístico N.° 114: Saber 11° 2024. Pontificia Universidad Javeriana. <https://www.javeriana.edu.co/recursosdb/d/lee/inf-114-informe-saber-11-2025-lee>

Ministerio de Educación de Argentina. (2022). Resultados de la evaluación Aprender. Secretaría de Evaluación Educativa.

Parsons, T. (2013). The social system. Routledge.

Razali, N. M., y Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1), 21–33.

Rendón-Macías, M. E., Villasís-Keever, M. Á., y Miranda-Navales, M. G. (2021). Métodos estadísticos para el análisis del tamaño del efecto. *Revista Alergia México*, 68(2), 128–136. https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2448-91902021000200128

UNESCO. (2022). Combatir los prejuicios y estereotipos de género en la educación y mediante esta. <https://www.unesco.org/es/articles/combatar-los-prejuicios-y-estereotipos-de-genero-en-la-educacion-y-mediante-esta>

UNESCO. (2024). PISA 2022: el panorama de los países de América Latina y el Caribe. Oficina Regional de Educación para América Latina y el Caribe. <https://www.unesco.org/es/articles/pisa-2022-el-panorama-de-los-paises-de-america-latina-y-el-caribe>

UNINORTE. (2025). Estereotipos de género en universitarias con alto rendimiento [Tesis de maestría, Universidad del Norte]. <https://manglar.uninorte.edu.co/>

Universidad de Chile. (2023). Estudio reveló que mujeres tienen menor rendimiento en pruebas de alta consecuencia debido a estereotipos. <https://uchile.cl/noticias>

Wasserstein, R. L., y Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, 70(2), 129–133.