

Bogotá D.C
Julio 2026
ISSN: 2590 - 4663
Publicación Trimestral



SABER al detalle

Edición 22

¿Qué hace comparables los
resultados de los exámenes Saber?



Saber Al Detalle: ¿Qué hace comparables los resultados de los exámenes Saber?

Subdirección de Estadísticas

Presidente de la República
Abelardo de la Espriella Otero

Ministra de Educación Nacional
Ilva Myriam Hoyos Castañeda

Directora General Icfes
Elizabeth Blandón Bermúdez

Director Técnico de Evaluación
Gustavo Andrés Monsalve Londoño

Subdirector de Diseño de Instrumentos
Heider Martínez Mena

Subdirector de Estadísticas
Cristian Fabián Montaña Rincón

Subdirectora de Análisis y Divulgación
Alejandra Neira Aroca

Elaboración del documento
Martha Ligia Cuevas Mendoza
John Alexander Calderón Rodríguez

Revisado por
Cristian Fabián Montaña Rincón

Diseño y diagramación
Linda Nathaly Sarmiento Olaya

Bogotá D.C., septiembre de 2026

Todos los derechos de autor reservados ©.

¿Qué hace comparables los resultados de los exámenes Saber?

Las evaluaciones estandarizadas a gran escala constituyen una herramienta fundamental para los sistemas educativos. Por una parte, permiten obtener información sobre el nivel de desarrollo de las habilidades y competencias de los evaluados y, por otra, proporcionan información para el monitoreo de la calidad educativa en diferentes niveles de agregación, tales como instituciones educativas, programas académicos, entidades territoriales y país. En Colombia, la evaluación estandarizada se efectúa a través de exámenes como Saber 3°, 5°, 7° y 9°, Saber 11°, Saber TyT y Saber Pro, desarrollados por el Icfes como parte del sistema nacional de evaluación de la educación.

La utilidad de estas evaluaciones depende, en gran medida, de la posibilidad de realizar comparaciones válidas entre los resultados obtenidos. La comparabilidad de los puntajes permite interpretar diferencias entre evaluados, grupos o periodos de aplicación, así como realizar seguimiento a la evolución de los resultados a lo largo del tiempo. Sin embargo, la comparabilidad no depende únicamente de que los puntajes se expresen en una misma escala, sino también de que las evaluaciones mantengan condiciones de medición equivalentes respecto al constructo evaluado, las especificaciones, el modelo de calificación, la población de referencia, la escala de reporte y otros componentes fundamentales de la evaluación.

Por esta razón, los procesos de diseño, construcción, aplicación, calificación,

equiparación y reporte de resultados se desarrollan bajo marcos técnicos que buscan garantizar la estabilidad de dichos componentes y, en consecuencia, la comparabilidad de los puntajes entre aplicaciones. Estos procesos quedan registrados en la normativa y los documentos técnicos que regulan los exámenes Saber que establecen estructuras de evaluación, marcos de referencia, metodologías de calificación y procedimientos para la producción de resultados comparables.

En este contexto, el Icfes mantiene la estabilidad de los componentes fundamentales del examen y emplea procedimientos de equiparación que permiten vincular diferentes aplicaciones a una misma escala de medición. No obstante, cuando se producen cambios en componentes esenciales de la evaluación, la comparabilidad puede verse afectada, por lo que resulta necesario analizar las condiciones bajo las cuales las comparaciones entre resultados pueden considerarse válidas.

Componentes de los exámenes estandarizados a gran escala

Como se mencionó anteriormente, en las evaluaciones estandarizadas a gran escala, la validez de las comparaciones entre resultados depende de la estabilidad de diversos componentes de la evaluación. Algunos de estos componentes se relacionan con el constructo y contenido evaluado, por ejemplo, las competencias evaluadas en

Saber 11.º, los módulos de Saber Pro y las especificaciones de estos. Otros corresponden a las propiedades de medición, entre ellos el modelo de calificación, la escala de reporte, la línea base y los procedimientos de equiparación. Por otra parte, la categoría de interpretación y el uso de las puntuaciones se encuentra asociada al reporte de resultados, mientras que la población evaluada y las condiciones de aplicación y administración comprenden un aspecto distinto.

De manera general, estas categorías permiten responder cuatro preguntas principales para establecer la comparabilidad de los resultados: si se sigue midiendo lo mismo, si se sigue midiendo de la misma manera, si las puntuaciones continúan interpretándose de forma equivalente y si se sigue evaluando a la misma población bajo condiciones comparables. Cuando se producen cambios en alguno de estos aspectos, la comparabilidad de los resultados puede verse afectada y las diferencias observadas podrían reflejar modificaciones en las características de la evaluación y no necesariamente en cambios reales en el desempeño de los evaluados.

Antes de analizar estas categorías en relación con las pruebas estandarizadas que aplica el Icfes, se presenta un resumen de los principales componentes que conforman la estructura de los exámenes estandarizados a gran escala. Estos elementos se encuentran definidos en los marcos de referencia de evaluaciones nacionales, como las desarrolladas por el Icfes, y en estudios internacionales como PISA, TIMSS y ERCE. Para cada componente se describe su función dentro de la evaluación y el efecto que sus modificaciones pueden tener sobre la comparabilidad de los resultados.

Dominios o constructos evaluados.

Corresponden a los conceptos, atributos, competencias, conocimientos, habilidades, capacidades, procesos, comportamientos o contenidos que una evaluación busca representar y sobre los cuales pretende sustentar interpretaciones a partir de los resultados. Aunque los términos dominio y constructo están estrechamente relacionados y su uso no es completamente uniforme en la literatura, es posible establecer una distinción analítica entre ellos. De acuerdo con AERA, APA, & NCME (2014), el constructo hace referencia al concepto, atributo, competencia o característica que una evaluación pretende medir y sobre el cual se busca realizar inferencias a partir de las puntuaciones, mientras que el dominio delimita el campo de conocimientos, habilidades, capacidades, comportamientos, contenidos o prácticas relevantes para la evaluación.

El constructo o dominio se representa en la prueba mediante especificaciones que determinan, entre otros aspectos, los contenidos, habilidades, procesos, contextos y características de las tareas que proporcionarán la evidencia para sustentar las inferencias previstas. En evaluaciones a gran escala como Saber 11º, Saber TyT y Saber Pro del Icfes, y en evaluaciones internacionales como PISA, TIMSS y ERCE, el objeto de evaluación se delimita y operacionaliza mediante marcos conceptuales y especificaciones propios de cada evaluación, que pueden experimentar modificaciones entre ciclos. Por ejemplo, en Saber 11º se evalúan Lectura Crítica, Matemáticas, Ciencias Naturales, Sociales y Ciudadanas e Inglés, mientras que en PISA se mantienen los dominios centrales de Lectura, Matemáticas y Ciencias, con rotación del dominio principal entre ciclos.

Para efectos de la comparabilidad longitudinal, resulta relevante distinguir entre cambios en la definición del constructo o dominio y cambios en la manera como estos se representan en la evaluación, por ejemplo, mediante la selección de contenidos, procesos, contextos o tipos de tareas. Una modificación en estos componentes no implica necesariamente un cambio en aquello que se pretende medir ni una ruptura de la comparabilidad, siempre que se preserve suficientemente el significado del constructo o dominio y la interpretación de los puntajes. Por el contrario, cuando las modificaciones alteran de manera sustantiva el objeto de medición o su representación, de modo que cambia el significado de las puntuaciones, pueden limitar la posibilidad de realizar comparaciones válidas a través del tiempo.

Especificaciones. Corresponden a los marcos de referencia que definen las competencias, procesos cognitivos, contenidos y contextos evaluados dentro de cada constructo o dominio. Estas especificaciones orientan la construcción de los ítems y delimitan lo que las pruebas pretenden medir. En el caso de los exámenes Saber del Icfes, se concretan en documentos técnicos basados en afirmaciones, evidencias y tareas evaluativas, mientras que en evaluaciones como PISA o TIMSS se formalizan mediante marcos de evaluación internacionales. Estas especificaciones fundamentan lo que se va a medir y hacen parte de la representación del constructo, por lo que cambios en su definición pueden afectar la forma en que este es operacionalizado y, en consecuencia, la comparabilidad de los resultados.

Modelo de calificación. Está asociado al modelo psicométrico empleado para la obtención de los puntajes. Entre los enfoques más utilizados se encuentran la Teoría Clásica

de los Tests (TCT) y la Teoría de Respuesta al Ítem (TRI). Mientras la TCT basa el puntaje en el número o porcentaje de respuestas correctas, la TRI estima la habilidad de los evaluados considerando las características psicométricas de los ítems. La utilización de modelos diferentes, incluso dentro del mismo enfoque, puede producir métricas distintas y afectar la comparabilidad de los resultados.

Escala de reporte. Corresponde a la métrica en la cual se expresan los resultados de la evaluación. Dependiendo del examen, los puntajes pueden reportarse mediante diferentes rangos, transformaciones o escalas estandarizadas. Aunque dos escalas puedan medir el mismo constructo, cambios en la métrica de reporte pueden dificultar la comparación directa de los resultados e inducir interpretaciones erróneas.

Línea base. Corresponde al periodo o cohorte de referencia con el cual se fija la escala de medición del examen. Usualmente, en una aplicación inicial se define una distribución de referencia para la población evaluada y, a partir de ella, se interpretan los resultados de las aplicaciones posteriores. Modificaciones en la línea base pueden alterar la interpretación de los puntajes aun cuando el desempeño de los evaluados permanezca estable.

Equiparación. Corresponde al conjunto de procedimientos estadísticos y psicométricos utilizados para hacer comparables distintas aplicaciones o formas de un examen en una misma escala de medición. En las pruebas del Icfes estos procedimientos se apoyan en modelos TRI y en el uso de ítems comunes entre aplicaciones. La equiparación constituye uno de los mecanismos fundamentales para garantizar la comparabilidad en el tiempo (longitudinal) de los resultados.

Reporte de resultados. Corresponde a las estadísticas e indicadores que se comunican a los evaluados en el examen. Estos pueden incluir puntajes, niveles de desempeño, percentiles, percentiles por subgrupos, puestos, desviaciones estándar e indicadores agregados. Cambios en los indicadores reportados o en su forma de cálculo pueden modificar la interpretación de los resultados y limitar las comparaciones entre aplicaciones.

Población evaluada. Corresponde al conjunto de personas sobre las cuales se pretende realizar inferencias a partir de los resultados de la evaluación. Incluye aspectos como la cobertura, los criterios de elegibilidad y exclusión, así como la decisión de realizar aplicaciones censales o muestrales. Cambios en la población objetivo pueden generar diferencias en los resultados que no necesariamente reflejan cambios reales en el desempeño.

Condiciones de aplicación. Corresponden al conjunto de condiciones operativas bajo las cuales se desarrolla la evaluación. Incluyen aspectos como los tiempos de aplicación, la distribución de las sesiones (por ejemplo, en uno o dos días), la modalidad de aplicación (papel o computador) y las condiciones logísticas de administración del examen. Modificaciones en estas condiciones pueden afectar el desempeño observado de los evaluados y, por tanto, la comparabilidad de los resultados entre aplicaciones.

Comparabilidad de puntajes mediante equiparación

Como se mencionó anteriormente, existen varios elementos que hacen parte de la estructura de la evaluación estandarizada.

Bajo el supuesto de que estos componentes se mantienen estables entre aplicaciones, la metodología mediante la cual se garantiza la comparabilidad de los puntajes entre distintas aplicaciones se denomina equiparación, como se señaló anteriormente.

La equiparación (equating) es el proceso mediante el cual los puntajes obtenidos en diferentes aplicaciones de un examen se transforman a una misma escala, de modo que sean directamente comparables. Este proceso busca asegurar que las diferencias en los resultados reflejen cambios reales en el desempeño de los evaluados y no variaciones en la dificultad de las pruebas. En la práctica, la equiparación corrige posibles diferencias entre formas de examen, permitiendo mantener la coherencia de la escala definida en la línea base y las propiedades de medición de la evaluación a lo largo del tiempo.

Existen diversas metodologías de equiparación, tanto en el marco de la TCT como de la TRI, incluyendo métodos lineales, equipercentiles y basados en modelos. Por ejemplo, entre los enfoques basados en la TRI se destaca el método de Fijación de Parámetros de Ítems Comunes (FCIP, por sus siglas en inglés, Fixed Common Item Parameters), que utiliza ítems compartidos para vincular distintas formas de prueba en una misma escala. En este método, los parámetros de dichos ítems se mantienen fijos en la escala definida, mientras que los parámetros de los nuevos ítems y las habilidades de los evaluados se estiman en esa misma escala. Este procedimiento permite contar con distintas formas o aplicaciones de una prueba en una única métrica, de manera que los puntajes reportados sean comparables.

En consecuencia, la equiparación expresa los puntajes de distintas aplicaciones en una misma escala, pero no garantiza por sí sola su comparabilidad. Esta requiere, además, que los principales componentes de la evaluación se mantengan estables entre aplicaciones. En el caso de las pruebas del Icfes, la metodología de equiparación empleada depende de las características de la evaluación y de los objetivos de comparabilidad establecidos.

Otros tipos de comparación y escalamiento

En las evaluaciones estandarizadas a gran escala como los exámenes Saber, el principal interés es realizar comparaciones válidas entre las poblaciones evaluadas a lo largo del tiempo; por esta razón, la equiparación constituye la metodología más adecuada. Sin embargo, en otros contextos de evaluación existen procedimientos psicométricos que permiten establecer distintos grados de relación entre los resultados obtenidos o construir escalas comunes de medición.

Una de estas metodologías es la vinculación de escalas (linking), en la cual, si se dispone de dos exámenes que miden constructos similares o relacionados, pero que no necesariamente cumplen los requisitos estrictos para ser considerados intercambiables, se establece una relación estadística entre sus escalas de puntuación. El propósito del linking es facilitar interpretaciones conjuntas o comparaciones aproximadas entre los resultados de ambos instrumentos, preservando las diferencias inherentes entre ellos. A diferencia de la equiparación, el linking no busca garantizar que los puntajes sean equivalentes ni intercambiables, sino proporcionar una

correspondencia que permita interpretar los resultados dentro de un marco común.

En este sentido, enfoques como el “Rosetta Stone”, utilizado para establecer la concordancia entre ERCE (Estudio Regional Comparativo y Explicativo) aplicado en algunos países de Latinoamérica y evaluaciones como TIMSS y PIRLS, presenta un caso particular de estos procedimientos, en el cual se construye una correspondencia estadística entre escalas de distintos sistemas de evaluación a partir de la alineación de dominios evaluados y de la identificación de patrones comunes de desempeño (UNESCO Institute for Statistics, 2022). Este enfoque permite aproximar resultados entre estudios o pruebas que no son directamente comparables, generando una base de referencia compartida para la interpretación de resultados a nivel internacional, sin implicar equivalencia estricta entre las pruebas.

Efecto de los cambios sobre la comparabilidad de los resultados

En la **Tabla 1** se presenta una síntesis de los principales componentes de las evaluaciones estandarizadas a gran escala y del efecto que tendría su modificación sobre la comparabilidad de los resultados. Para cada elemento se describe el tipo de efecto que puede generarse ante un cambio en sus características y la forma en que este es conceptualizado en la literatura especializada. Esta clasificación permite identificar de manera estructurada si las diferencias observadas entre aplicaciones responden a variaciones reales en el desempeño de los evaluados o a modificaciones en el constructo evaluado, las propiedades de medición, la interpretación de los resultados o la población y condiciones de aplicación.

• **Tabla 1. Componentes de la evaluación y su efecto sobre la comparabilidad de los resultados**

Aspecto	Componente	Si cambia, ¿qué ocurre? (Impacto en comparabilidad)	Tipo de afectación
Constructo y contenido	Constructos o dominios evaluados	Se modifica el constructo o el dominio evaluado (por ejemplo, incluir o excluir áreas o cambiar su peso relativo). Los resultados dejan de representar el mismo constructo global.	Cambio en el constructo o en el dominio.
	Especificaciones	Cambian las competencias, contenidos, procesos cognitivos o contextos evaluados. Aun cuando el nombre de la prueba se mantenga, puede cambiar lo que realmente se mide.	Cambio en el constructo o dominio y en las evidencias de validez relacionadas con el contenido de la prueba.
Propiedades de medición	Modelo de calificación	Cambia la forma de estimar la habilidad o el desempeño de los evaluados. Los puntajes obtenidos bajo diferentes modelos pueden no representar la misma métrica.	No hay equivalencia métrica.
	Escala de reporte	Se modifica la escala en la cual se expresan los resultados (por ejemplo, rangos de puntaje o métricas reportadas). Las diferencias observadas pueden deberse al cambio de escala y no al desempeño.	No hay equivalencia métrica.
	Línea base	Se redefine la población o aplicación de referencia utilizada para fijar la escala. Cambia la interpretación relativa de los puntajes.	Re-escalamiento.
	Equiparación	Cambian los procedimientos utilizados para equiparar las aplicaciones o formas del examen. Las diferencias observadas pueden reflejar variaciones metodológicas y no cambios reales en el desempeño.	Ruptura de la escala longitudinal.
Interpretación y uso	Reporte de resultados	Cambian los indicadores reportados o su forma de cálculo (puntajes, niveles de desempeño, percentiles, puestos, etc.). Puede generarse una interpretación diferente de resultados equivalentes.	Cambio en la interpretación de resultados.

Aspecto	Componente	Si cambia, ¿qué ocurre? (Impacto en comparabilidad)	Tipo de afectación
Población y condiciones de aplicación	Población evaluada	Cambia la población objetivo, la cobertura o los criterios de inclusión y exclusión. Los resultados dejan de representar el mismo universo de evaluados.	Cambio en la población objetivo.
	Condiciones de aplicación	Cambian aspectos como tiempos (un día o dos), modalidad (papel o computador), orden de presentación en las pruebas o condiciones operativas de la aplicación.	Cambio en las condiciones de medición.

Principales cambios en los Exámenes Saber

Teniendo en cuenta lo expuesto anteriormente, en este apartado se presenta una síntesis de la evolución histórica de los principales componentes de los exámenes Saber, según los cuatro aspectos que sustentan la comparabilidad de las evaluaciones estandarizadas a gran escala.

Saber 3°, 5°, 7° y 9°

La evolución de las pruebas Saber para los grados 3°, 5°, 7° y 9° puede entenderse a partir de cuatro grandes etapas. La primera comprende las evaluaciones nacionales muestrales desarrolladas durante la década de 1990 y la transición hacia las primeras aplicaciones censales iniciadas en 2002. Estas evaluaciones muestrales, concebidas para producir información representativa del sistema educativo, evolucionaron hacia un esquema censal impulsado por la Ley 715 de 2001, con el propósito de generar resultados para todos los establecimientos educativos y fortalecer los procesos de mejoramiento institucional.

Entre 2002 y 2008, se desarrolló la segunda etapa en la que estas pruebas experimentaron una fase de construcción y ajuste metodológico caracterizada por múltiples operativos con diferencias en los grados evaluados, las áreas incluidas, la estructura de los cuadernillos y los procedimientos de aplicación. Durante este período se consolidó el enfoque de evaluación por competencias, se incorporaron progresivamente áreas como Ciencias Sociales y Competencias Ciudadanas, se introdujo el diseño matricial de cuadernillos y se inició la transición hacia modelos psicométricos basados en la Teoría de Respuesta al Ítem (TRI). Sin embargo, la heterogeneidad de estas aplicaciones impidió conformar una serie histórica continua, situación que posteriormente motivó el desarrollo de un estudio de equiparación utilizando la aplicación de 2009 como referencia para expresar los resultados de 2002, 2005, 2006 y 2009 sobre una misma escala de medición (Icfes, 2011a).

La tercera etapa corresponde a la consolidación técnica del programa entre 2009 y 2017. Durante este

período se estabilizaron los constructos, las especificaciones de las pruebas, los procedimientos de ensamblaje de cuadernillos, el uso sistemático de la TRI, las escalas de reporte y los niveles de desempeño (Ver Icfes, 2010; Icfes, 2011b). Asimismo, se fortalecieron los estudios de factores asociados, se ampliaron los reportes para establecimientos educativos y entidades territoriales y, posteriormente, se incorporaron las afirmaciones como un nuevo nivel de reporte para apoyar la interpretación pedagógica de los resultados. Estas transformaciones permitieron consolidar una serie histórica sustentada en una misma línea de medición y fortalecer la comparabilidad longitudinal de las pruebas.

Finalmente, la cuarta etapa corresponde a la actualización conceptual implementada entre 2019 y 2024, que dio origen a una nueva generación de las pruebas. Esta transformación incluyó nuevos marcos de referencia, especificaciones, escalas de reporte, línea base, la incorporación oficial del

grado séptimo, la medición de habilidades socioemocionales y el retorno a un esquema de aplicación muestral para el seguimiento del sistema educativo. Debido a estos cambios conceptuales y metodológicos, el Icfes estableció nuevas líneas base a partir de 2022 para los grados 3°, 5° y 9°, y de 2023 para el grado 7°, iniciando una nueva serie histórica de resultados (Icfes, 2023).

En la **Tabla 2** se sintetizan las principales características de cada una de estas etapas a partir de los cuatro aspectos que sustentan la comparabilidad de las evaluaciones estandarizadas a gran escala: el constructo y contenido, las propiedades de medición, la interpretación y uso de los resultados y la población y condiciones de aplicación. Esta síntesis permite identificar la evolución técnica del examen y reconocer aquellos cambios que dieron lugar a nuevas líneas de medición o modificaron las condiciones necesarias para realizar comparaciones válidas entre los resultados de diferentes aplicaciones.

• **Tabla 2.** Principales características de las pruebas Saber 3°, 5°, 7° y 9° según los aspectos de la comparabilidad de los resultados

Aspecto	1990–2001	2002–2008	2009–2017	2022–Actualidad
Constructo y contenido	• Transición de contenidos a competencias. • Lenguaje, Matemáticas y otras áreas según el operativo.	• Consolidación del enfoque por competencias. • Incorporación de Ciencias Sociales y Competencias Ciudadanas. • Especificaciones en consolidación.	• Competencias y componentes estables. • Especificaciones consolidadas. • Base del Modelo Basado en Evidencias (MBE).	• Nuevos marcos de referencia. • Actualización de competencias, afirmaciones y evidencias. • Incorporación de habilidades sociales y emocionales.

Aspecto	1990–2001	2002–2008	2009–2017	2022–Actualidad
Propiedades de medición	• Puntajes observados. • Escalas propias de cada aplicación.	• Transición hacia modelos basados en TRI. • Diseño matricial de cuadernillos. • Primeros procesos de equiparación.	• TRI como modelo de calificación. • Ensamblaje estandarizado. • Equiparación y escala común.	• Continuidad de TRI. • Nuevas escalas de reporte (100–700 y 0–10 en Acciones y actitudes). • Nuevas líneas base (2022–2023).
Interpretación y uso	• Resultados agregados del sistema. • Uso para monitoreo nacional.	• Resultados por establecimiento educativo. • Fortalecimiento de los reportes institucionales.	• Puntajes y niveles de desempeño con metodologías definidas. • Reportes por establecimiento, Entidades Territoriales Certificadas y país. • Factores asociados y reportes por afirmaciones.	• Reportes alineados con los nuevos marcos. • Integración de habilidades socioemocionales y contexto. • Nueva línea de interpretación desde 2022.
Población y condiciones de aplicación	• Evaluaciones muestrales. • Inferencia sobre el sistema educativo.	• Aplicaciones censales no controladas (realizadas bajo la responsabilidad de la institución). • Operativos con diferencias entre aplicaciones.	• Aplicaciones censales estandarizadas no controladas (realizadas bajo la responsabilidad de la institución). • Desde 2014, articulación de operativos censales y muestrales.	• Retorno a evaluación muestral. • Incorporación del grado 7° de manera oficial. • Seguimiento del sistema educativo.

Saber 11

La evolución del examen Saber 11° puede entenderse a partir de tres grandes etapas. La primera corresponde al período comprendido entre 1968 y 1999, caracterizado por una evaluación centrada

inicialmente en aptitudes y conocimientos, con una finalidad principalmente selectiva para el ingreso a la educación superior. Durante esta etapa, la estructura, el número de áreas y los rangos de calificación cambiaron varias veces: entre 1968 y 1979 se evaluaron entre ocho y nueve áreas y el

puntaje total podía estar aproximadamente entre 180 y 640 puntos; posteriormente, entre 1981 y 1999, el puntaje total se ubicó alrededor de 100 y 400. Además, durante la década de 1990 el examen se organizó mediante un núcleo común y pruebas electivas, por lo que no todos los estudiantes respondían exactamente el mismo conjunto de pruebas (Ver Cardona, 2025; Grupo de Evaluación de la Educación Básica y Media, 1999; Grupo de Investigación sobre Pruebas Masivas en Colombia, 2008; Icfes, 2013; Viña, 2024). La calificación se fundamentaba principalmente en la TCT, lo que, sumado a los cambios estructurales y de escala, impediría considerar comparables las aplicaciones de este período entre sí o con las versiones posteriores.

La segunda etapa se inicia en el año 2000 con la transición hacia un enfoque basado en competencias. El examen pasó a evaluar la capacidad de interpretar, argumentar y resolver problemas, aunque mantuvo una estructura de ocho pruebas y un componente flexible (Viña, 2024). Entre 2000 y 2004, los puntajes por prueba se reportaron en una escala con media 50 y desviación estándar 10, con un mínimo de 0 y sin un máximo acotado (Oficina de Gestión de Proyectos de Investigación, 2025). A partir de 2005 se consolidó el uso de la TRI mediante el modelo de Rasch, con una escala igual a la del período inmediatamente anterior. Este cambio permitió estimar la habilidad con menor dependencia del conjunto específico de preguntas y fortalecer los procedimientos

de equiparación. Por ello, los resultados de las pruebas del examen obtenidos entre 2005 y el primer semestre de 2014 podrían considerarse comparables dentro de este período, aunque existen variaciones menores asociadas a cambios metodológicos y ajustes realizados por el Icfes.

La tercera etapa comenzó en el segundo semestre de 2014 con la alineación de Saber 11° dentro del Sistema Nacional de Evaluación Estandarizada. El examen se reorganizó en cinco pruebas: Lectura Crítica, Matemáticas, Sociales y Ciudadanas, Ciencias Naturales e Inglés, e incorporó puntajes por prueba de 0 a 100, un puntaje global de 0 a 500, percentiles y niveles de desempeño. Entre 2014-2 y 2016-1 se mantuvo transitoriamente el modelo de Rasch; posteriormente, los resultados fueron recalificados con el modelo logístico de tres parámetros (3PL), utilizado desde 2016-2. Asimismo, se equipararon y recalificaron algunas aplicaciones de 2012 a 2014-1, cuyos resultados son solo parcialmente comparables con la escala vigente (Acero et al., 2016). Desde 2014-2 se configura la serie histórica actual, cuyos resultados son comparables entre sí conforme a las reglas establecidas en la Resolución 268 de 2020 (Icfes, 2020).

En la **Tabla 3** se sintetizan las principales características de cada uno de estos períodos a partir de los cuatro aspectos que sustentan la comparabilidad de las evaluaciones estandarizadas a gran escala.

• **Tabla 3. Principales características del examen Saber 11° según los aspectos de la comparabilidad de los resultados.**

Aspecto	1968–1999	2000–2014-1	2014-2–Actualidad
Construto y contenido	• Aptitudes y conocimientos. • Núcleo común y pruebas electivas.	• Ocho pruebas y un componente flexible, • Enfoque por competencias. • Marcos de referencia. • Especificaciones por competencias.	• Cinco pruebas. • Especificaciones actualizadas. • Afirmaciones y evidencias. • Modelo basado en evidencias.
Propiedades de medición	• TCT. • Puntajes observados. • Escalas propias de cada período.	• Transición de TCT a TRI (Modelo de un parámetro - 1PL). • Equiparación. • Escala comparable.	• TRI (Modelo de tres parámetros - 3PL). • Línea base (media 50, desviación 10). • Equiparación. • Escala comparable.
Interpretación y uso	• Reportes de resultados propios del período. • Interpretación centrada en aptitudes y conocimientos.	• Fortalecimiento de los reportes e indicadores. • Mayor uso para seguimiento del sistema educativo y de las instituciones.	• Puntaje global y puntajes por prueba. • Percentiles y niveles de desempeño. • Reportes con mayor desagregación para distintos usuarios.
Población y condiciones de aplicación	• Grupos diferenciados de aspirantes. • Aplicación en papel.	• Población objetivo estable. • Aplicación estandarizada en papel. • Dos sesiones de aplicación a partir de 2005.	• Población objetivo estable. • Ajustes en condiciones de aplicación. • Aplicación en papel y computador para poblaciones específicas. • Dos sesiones de aplicación.

Saber Pro

La evolución del examen Saber Pro puede entenderse a partir de tres grandes etapas. La primera corresponde a los Exámenes de Calidad de la Educación Superior (ECAES),

implementados entre 2003 y 2009, los cuales fueron concebidos como evaluaciones específicas para cada programa académico con el propósito de apoyar los procesos de calidad de la educación superior (Presidencia de la República, 2003). Durante este período,

las pruebas fueron construidas con la participación de comunidades académicas y respondían a referentes disciplinares propios de cada profesión, por lo que diferían en sus constructos, especificaciones, estructuras, procedimientos de ensamblaje y modelos de medición. Aunque en los últimos años de este periodo comenzaron a incorporarse procedimientos basados en la TRI, la ausencia de competencias genéricas comunes y de una estructura estandarizada impidió conformar una serie histórica comparable entre programas y con las evaluaciones posteriores.

La segunda etapa inicia con la expedición de la Ley 1324 de 2009 y del Decreto 3963 de 2009, que transformaron los ECAES en el examen Saber Pro y redefinieron el sistema de evaluación de la educación superior en Colombia (Congreso de Colombia, 2009; Presidencia de la República, 2009). Este proceso estuvo precedido por el proyecto Graduate Skills Assessment (GSA), desarrollado conjuntamente por el Icfes y el Australian Council for Educational Research (ACER), mediante el cual se diseñó un modelo de evaluación basado en competencias genéricas comunes a todos los programas académicos (Icfes, 2011c). A partir de esta transformación se consolidó una estructura integrada por módulos genéricos y específicos, se estandarizaron las especificaciones técnicas (2011d), se fortaleció el uso de la TRI, los procesos de ensamblaje y los reportes de resultados, y se establecieron grupos de referencia

como unidad de comparación para los módulos específicos. Estas modificaciones representaron una ruptura metodológica respecto a los ECAES y dieron origen a una nueva serie histórica del examen.

La tercera etapa corresponde a la consolidación metodológica de Saber Pro entre 2016 y la actualidad. Durante este periodo se fortaleció el Modelo Basado en Evidencias, se consolidó el uso de la TRI mediante el modelo logístico de tres parámetros para los módulos genéricos, se implementó una nueva línea base con promedio de 150 puntos y desviación estándar de 30 puntos, se estandarizaron los procesos de equiparación, el armado de cuadernillos y los reportes institucionales, y se fortalecieron los niveles de desempeño como apoyo para la interpretación de los resultados. De acuerdo con la Resolución 268 de 2020, el Icfes garantiza la comparabilidad de los resultados obtenidos desde la línea base establecida en 2016 (Icfes, 2020); en consecuencia, los resultados de aplicaciones anteriores no son técnicamente comparables con los reportes actuales debido a los cambios en el constructo evaluado, las escalas de reporte, los modelos psicométricos y la estructura del examen.

En la **Tabla 4** se sintetizan las principales características de cada período a partir de los cuatro aspectos que sustentan la comparabilidad de las evaluaciones estandarizadas a gran escala.

• **Tabla 4. Principales características de los exámenes Saber Pro según los aspectos de la comparabilidad de los resultados**

Aspecto	ECAES (2003–2009)	Transición a Saber Pro (2009–2015)	Saber Pro (2016–Actualidad)
Construido y contenido	• Competencias específicas por profesión. • Marcos y especificaciones independientes por programa.	• Competencias genéricas y específicas. • Estructura modular. • Especificaciones comunes para módulos genéricos. • Modelo Basado en Evidencias.	• Competencias genéricas y específicas. • Marcos de referencia, afirmaciones y evidencias estandarizadas.
Propiedades de medición	• TCT y TRI (Rasch) para análisis psicométrico. • Sin escalas comunes entre programas o aplicaciones.	• TRI como base de la calificación. • Escalas comunes para competencias genéricas. • Ensamblaje modular de cuadernillos.	• TRI para estimación y reporte. • Ensamblaje provisto con especificaciones y parámetros psicométricos. • Comparabilidad mediante bancos calibrados.
Interpretación y uso	• Resultados orientados a programas e instituciones. • Énfasis en aseguramiento de la calidad.	• Reportes para estudiantes, programas e instituciones. • Comparación por competencias genéricas y grupos de referencia.	• Reportes para estudiantes, programas, instituciones, Núcleos Básicos de Conocimiento y sistema nacional. • Percentiles, comparaciones con grupos de referencia e indicadores de valor agregado.
Población y condiciones de aplicación	• Estudiantes próximos a graduarse de programas específicos. • Aplicaciones independientes por profesión.	• Estudiantes próximos a culminar programas de educación superior. • Módulos genéricos comunes y específicos por grupo de referencia.	• Procedimientos de aplicación estandarizados para cada población. • Aplicación en papel y computador para población en el exterior y condiciones específicas de aplicación (por ejemplo, pandemia).

Saber TyT

La evolución del examen Saber TyT puede entenderse a partir de tres grandes momentos. El primero corresponde al período comprendido entre 2003 y 2010, cuando algunos programas técnicos profesionales y tecnológicos participaron en los Exámenes de Calidad de la Educación Superior ECAES. Entre 2003 y 2009 se aplicaron 55 exámenes para programas de educación superior, de los cuales únicamente tres estuvieron dirigidos a programas técnicos y tecnológicos (Administración y afines, Electrónica y afines y Sistemas y afines) (Icfes, 2017). Estas evaluaciones fueron construidas bajo una lógica disciplinar, con especificaciones independientes para cada programa académico y sin un marco conceptual común para este nivel de formación, por lo que sus resultados constituyen un antecedente histórico, pero no una serie comparable con el examen actual.

El segundo momento comprende el período 2011-2015, durante el cual el Icfes desarrolló, en conjunto con la Red de Instituciones Técnicas Profesionales, Tecnológicas y Universitarias Públicas (REDTTU), el proceso de diseño conceptual y metodológico del nuevo examen. En esta fase se construyeron los marcos de referencia, las especificaciones de prueba, los bancos de ítems y los grupos de referencia. Estas actividades permitieron pasar de una evaluación centrada en contenidos disciplinares a un modelo basado en competencias genéricas y específicas, sustentado en el Modelo Basado en Evidencias y en la TRI. Aunque aún no existía

una aplicación operativa, durante este período se establecieron las condiciones técnicas necesarias para garantizar la comparabilidad de las futuras aplicaciones (Icfes, 2017).

La tercera etapa inicia con la primera aplicación oficial de Saber TyT en 2016, que constituye la línea base de la serie histórica vigente. Desde entonces el examen ha mantenido una estructura modular compuesta por competencias genéricas y módulos específicos organizados por grupos de referencia, utilizando procedimientos estandarizados de construcción, equiparación y calificación fundamentados en la TRI. Entre 2016 y la actualidad se consolidaron los bancos de ítems, los reportes institucionales, los niveles de desempeño y el uso del MBE como fundamento conceptual del examen. Durante la pandemia por COVID-19 se incorporó progresivamente la aplicación por computador, y los estudios técnicos del Icfes evidenciaron que esta transición no afectó la comparabilidad de las escalas ni la interpretación de los resultados referencia (Icfes, 2021). En consecuencia, la serie histórica de Saber TyT puede considerarse técnicamente comparable desde 2016, siempre que las interpretaciones tengan en cuenta el contexto excepcional de las aplicaciones realizadas durante la pandemia y las actualizaciones menores efectuadas en los marcos de referencia.

En la **Tabla 5** se sintetizan las principales características de cada período a partir de los cuatro aspectos que sustentan la comparabilidad de las evaluaciones estandarizadas a gran escala.

• **Tabla 5. Principales características de los exámenes Saber TyT según los aspectos de la comparabilidad de los resultados.**

Aspecto	ECAES para programas técnicos y tecnológicos (2003–2010)	Construcción de Saber TyT (2011–2015)	Saber TyT (2016–Actualidad)
Construido y contenido	• Competencias específicas definidas para algunos programas técnicos y tecnológicos. • Marcos y especificaciones independientes para cada programa. • Evaluación con enfoque predominantemente disciplinar.	• Definición de competencias genéricas y específicas para la formación técnica y tecnológica. • Construcción de marcos de referencia con la REDTTU. • Adopción del Modelo Basado en Evidencias. • Estructura modular por competencias.	• Competencias genéricas y módulos específicos por grupos de referencia. • Marcos de referencia, afirmaciones y evidencias estandarizadas.
Propiedades de medición	• Uso de TCT y posteriormente TRI (Rasch) en algunos exámenes. • Sin escalas comunes entre programas. • Comparabilidad limitada a cada examen específico.	• Calibración de bancos de ítems. • TRI como base de la medición. • Definición de escalas y procedimientos de equiparación para la futura línea base.	• TRI para estimación y reporte. • Ensamblaje provisto con especificaciones y parámetros psicométricos. • Comparabilidad mediante bancos calibrados.
Interpretación y uso	• Resultados orientados al aseguramiento de la calidad de programas específicos. • Sin comparación transversal entre diferentes áreas de formación.	• Definición de reportes para estudiantes e instituciones. • Construcción de grupos de referencia para módulos específicos. • Diseño de niveles de desempeño.	• Reportes para estudiantes, programas, instituciones, Núcleos Básicos de Conocimiento y sistema nacional. • Percentiles, comparaciones con grupos de referencia e indicadores de valor agregado.
Población y condiciones de aplicación	• Estudiantes de último año de los programas académicos. • Exámenes específicos para cada programa académico. • Sin módulos genéricos comunes para todos los evaluados.	• Estudiantes próximos a culminar programas técnicos profesionales y tecnológicos. • Módulos genéricos comunes y específicos por grupo de referencia.	• Estudiantes próximos a culminar programas técnicos profesionales y tecnológicos. • Procedimientos de aplicación estandarizados para cada población. • Aplicación en papel y computador para población en el exterior y condiciones específicas de aplicación (por ejemplo, pandemia).

Comparaciones inadecuadas

En este documento se han descrito los principales componentes que conforman la estructura de una evaluación estandarizada a gran escala y cómo las diferencias en estos elementos pueden afectar la comparabilidad de los resultados. Por lo tanto, para que dos puntuaciones puedan compararse de manera válida, es necesario que provengan de evaluaciones que midan el mismo constructo, y utilicen modelos de calificación y escalas equivalentes. Cuando estas condiciones no se cumplen, las diferencias observadas pueden obedecer a cambios en las características de la medición y no necesariamente a diferencias reales en el desempeño de los evaluados.

A continuación, se presentan algunos ejemplos de comparaciones que suelen realizarse con frecuencia, pero que no resultan adecuadas debido a las diferencias existentes entre los componentes de las evaluaciones de donde provienen las puntuaciones.

Una comparación frecuente consiste en contrastar las calificaciones obtenidas por los estudiantes en asignaturas de la educación media o superior con los resultados alcanzados en las pruebas o módulos de los exámenes Saber 11, Saber TyT o Saber Pro que guardan alguna relación temática con dichas asignaturas. No obstante, esta comparación resulta inadecuada debido a que las calificaciones académicas responden a criterios de evaluación definidos por cada institución educativa o docente, los cuales pueden variar significativamente entre establecimientos, programas académicos y periodos, e incorporar elementos de valoración subjetiva. En consecuencia, las

puntuaciones obtenidas en ambos procesos de evaluación no son comparables.

Otra comparación frecuente consiste en contrastar los resultados obtenidos en simulacros, pruebas preparatorias o exámenes de entrenamiento con los puntajes reportados en los exámenes oficiales Saber. No obstante, esta comparación resulta inadecuada debido a que estos instrumentos suelen presentar diferencias en las especificaciones, la construcción de los ítems, la longitud de la prueba, el modelo de calificación y la población utilizada para su calibración. Adicionalmente, los simulacros generalmente no se encuentran vinculados a la misma línea base ni a los procedimientos de equiparación empleados en los exámenes oficiales, por lo que sus resultados se expresan en escalas diferentes. Por consiguiente, aunque pueda existir alguna relación entre el desempeño observado en un simulacro y el obtenido posteriormente en un examen oficial, las puntuaciones obtenidas en ambos procesos de evaluación no son comparables de ninguna manera.

Conclusiones

A lo largo de este documento se revisaron los principales componentes que conforman la estructura de una evaluación estandarizada, incluyendo las pruebas o módulos evaluados, las especificaciones, el modelo de calificación, la escala de reporte, el método de equiparación, la línea base, el reporte de resultados, la población evaluada y las condiciones de aplicación. La estabilidad de estos componentes resulta necesaria para garantizar la comparabilidad de los puntajes entre distintas aplicaciones.

Los exámenes Saber han evolucionado a lo largo del tiempo e incorporado cambios en su diseño y estructura.

Algunos de estos cambios han sido lo suficientemente relevantes como para limitar la comparación directa de los resultados entre determinados periodos de aplicación. En otros casos, el Icfes ha implementado procedimientos psicométricos y estadísticos que permiten mantener la comparabilidad de los puntajes frente a modificaciones menores en la evaluación.

En términos generales, las comparaciones válidas requieren que las puntuaciones provengan de evaluaciones con condiciones de medición equivalentes. Cuando existen diferencias en el constructo evaluado, las especificaciones, las propiedades de medición, la escala o la población de referencia, las puntuaciones dejan de ser directamente comparables. En estos casos, las diferencias observadas pueden reflejar cambios en las características de la medición y no necesariamente en el desempeño de los evaluados, lo que puede conducir a interpretaciones erróneas.

Referencias

Acero, W. F., Sánchez, J. F., Suárez, D., & Téllez, C. (2016). Modelo de recalificación para la prueba Saber 11. *Comunicaciones en Estadística*, 9(1), 43–54.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Cardona, C. A. (2025). Punto de cambio: De los contenidos a las competencias en la prueba Saber 11 en Colombia. *Dialéctica*.

Revista de Investigación Educativa, 2(27), 27–58. <https://doi.org/10.56219/dialctica.v2i27.5224>

Congreso de Colombia. (2009). *Ley 1324 de 2009, por la cual se fijan parámetros y criterios para organizar el sistema de evaluación de resultados de la calidad de la educación*.

Grupo de Evaluación de la Educación Básica y Media. (1999). *Antecedentes y marco legal del Examen de Estado*. Instituto Colombiano para el Fomento de la Educación Superior (ICFES).

- Grupo de Investigación sobre Pruebas Masivas en Colombia. (2008). El surgimiento del Servicio Nacional de Pruebas del ICFES en las voces de sus protagonistas. *Magistro*, 2(3), 115–134.
- Instituto Colombiano para la Evaluación de la Educación (Icfes). (2010). *SABER 5° y 9° 2009. Resultados nacionales: resumen ejecutivo*. Icfes.
- Instituto Colombiano para la Evaluación de la Educación (Icfes). (2011a). *Saber 5° y 9°. Informe técnico de resultados históricos*. Icfes.
- Instituto Colombiano para la Evaluación de la Educación (Icfes). (2011b). *Informe técnico. Saber 5° y 9°*. Icfes.
- Instituto Colombiano para la Evaluación de la Educación (Icfes). (2011c). *Prueba de habilidades genéricas GSA Colombia. Resultados del pilotaje*. Icfes.
- Instituto Colombiano para la Evaluación de la Educación (Icfes). (2011d). *Informe de gestión 2010*. Icfes.
- Instituto Colombiano para la Evaluación de la Educación. (2013). *Sistema Nacional de Evaluación Estandarizada de la Educación: alineación del examen Saber 11°*. Icfes.
- Instituto Colombiano para la Evaluación de la Educación (Icfes). (2017). *Saber TyT. Informe nacional de resultados. Línea de base 2016*.
- Instituto Colombiano para la Evaluación de la Educación (Icfes). (2020). *Resolución* 268 de 2020, por la cual se establecen las reglas para la calificación de los exámenes de Estado. Icfes.
- Instituto Colombiano para la Evaluación de la Educación (Icfes). (2021). *Informe técnico sobre el examen de Estado en tiempos de la COVID-19. Saber TyT y Saber Pro*.
- Instituto Colombiano para la Evaluación de la Educación (Icfes). (2023). *Informe nacional de resultados de las pruebas Saber 3°, 5°, 7° y 9° 2023*. Icfes.
- Oficina de Gestión de Proyectos de Investigación. (2025). *Guía metodológica para el acceso y uso de la información: Examen Saber 11°*. Instituto Colombiano para la Evaluación de la Educación (Icfes).
- Presidencia de la República de Colombia. (2003). *Decreto 1781 de 2003, por el cual se reglamentan los Exámenes de Estado de Calidad de la Educación Superior (ECAES)*.
- Presidencia de la República de Colombia. (2009). *Decreto 3963 de 2009, por el cual se reglamenta el Examen de Estado de Calidad de la Educación Superior*.
- UNESCO Institute for Statistics. (2022). *Rosetta Stone analysis report: Establishing a concordance between ERCE and TIMSS/PIRLS*. UIS.
- Viña, O. (2024). Las pruebas de Estado en Colombia para estudiantes de undécimo grado: una revisión histórica. *Línea Imaginaria*, 19, 533–550.