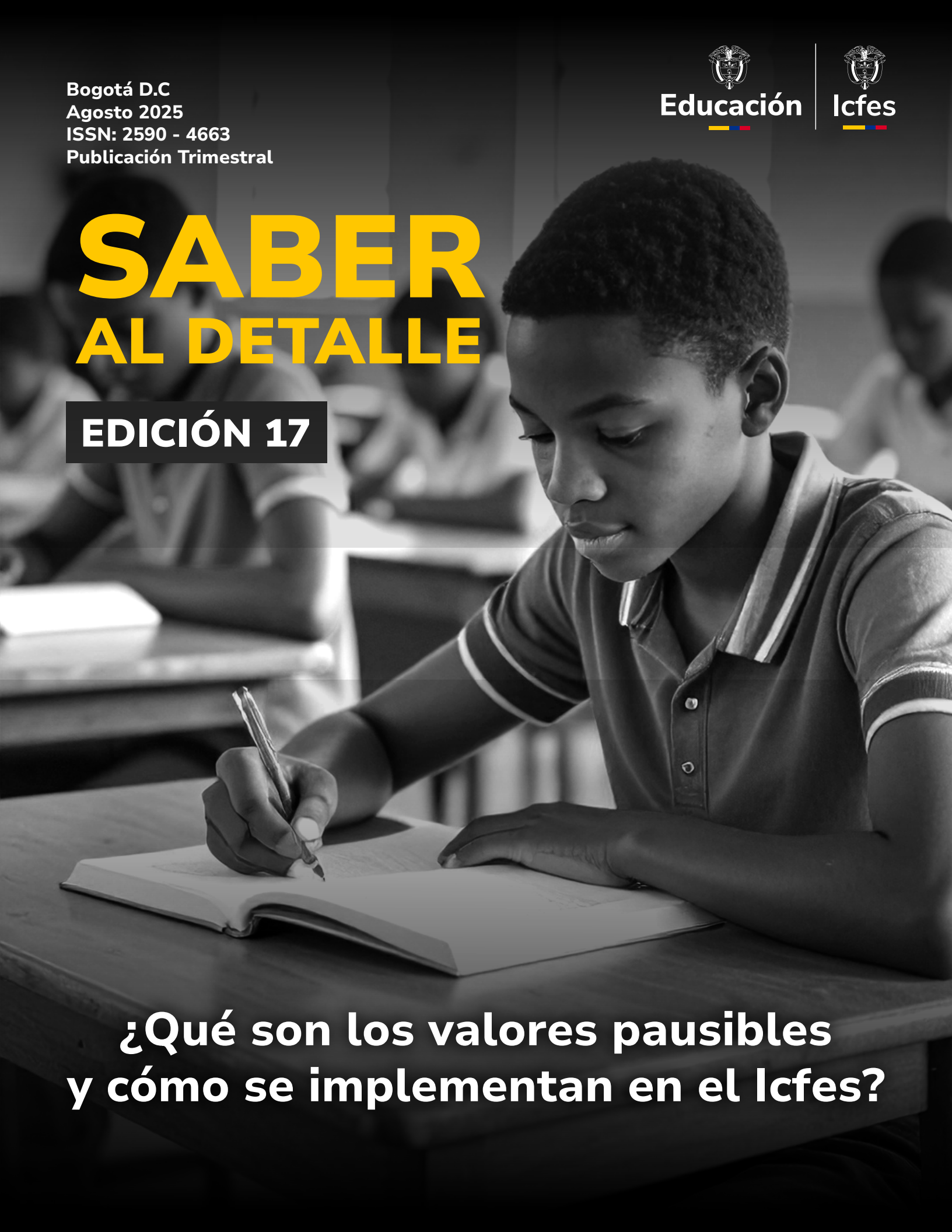
Bogotá D.C
Agosto 2025
ISSN: 2590 - 4663
Publicación Trimestral

Educación | Icfes

SABER AL DETALLE

EDICIÓN 17



¿Qué son los valores pausibles y cómo se implementan en el Icfes?

Elaboración del documento
John Alexander Calderón Rodríguez
Oscar Alejandro Angarita Rodríguez
Juan Felipe Salamanca Garzón

Revisado por
Ana María Mondragón Moreno
Juan Carlos Rubriche Cárdenas

Diseño y diagramación
Julián Andrés Castro Salvador

Bogotá D.C., Agosto 2025

¿Qué son los valores plausibles y cómo se implementan en el Icfes?

Las evaluaciones estandarizadas a gran escala son estudios diseñados para medir las competencias de los estudiantes en determinados dominios, como lectura, matemáticas, entre otras. En muchas de estas el objetivo principal no es evaluar a cada individuo, sino obtener información confiable sobre el desempeño de grupos poblacionales, como países, regiones o subgrupos específicos. Entre las evaluaciones internacionales más conocidas se encuentran PISA[1], TIMSS[2] y ERCE[3] y en las nacionales la prueba Saber 3°, 5°, 7° y 9°, recogen información relevante a nivel internacional y nacional permitiendo orientar las políticas educativas y promoviendo mejoras en la calidad de la educación.

Estas evaluaciones suelen abarcar un amplio rango de competencias cognitivas, lo que requiere incluir múltiples áreas y temas de conocimiento. Por ejemplo, PISA evalúa lectura, matemáticas y ciencias, lo que implica un conjunto extenso de contenidos. Para lograr una evaluación adecuada de estas áreas, sería necesario aplicar una gran cantidad de preguntas, lo cual resultaría poco factible si cada estudiante tuviera que responder a todos los ítems disponibles. Esta limitación metodológica es común en las evaluaciones a gran escala, ya que aplicar a un estudiante el total de ítems implicaría una alta demanda de tiempo y una carga cognitiva difícil de gestionar en la evaluación.

Por lo anterior, es común que en estas evaluaciones se utilice un diseño por bloques[5], en el que cada participante responde únicamente a un subconjunto del total de preguntas. Este enfoque permite evaluar un currículo amplio sin sobrecargar al estudiante ni prolongar excesivamente la duración de la prueba. No obstante, dicho diseño implica que cada estudiante tenga datos faltantes en una proporción considerable de los ítems, lo que representa un desafío metodológico para estimar con precisión su nivel de competencia. Para superar esta limitación, se recurre al uso de valores plausibles, una técnica que permite generar estimaciones válidas del desempeño a partir de la información disponible.

¿Qué son los valores plausibles y cuál es su utilidad?

Los valores plausibles se introducen como una solución técnica para realizar análisis estadísticos válidos a pesar de que los estudiantes no sean evaluados con todos los ítems. Dado que no se cuenta con suficiente información por estudiante para hacer estimaciones precisas de su nivel de competencia, los valores plausibles permiten imputar múltiples estimaciones de habilidad por evaluado a partir de sus respuestas y características contextuales, simulando distintas posibilidades dentro de su distribución posterior de competencia. Esto permite estimar estadísticas como medias, varianzas y correlaciones con mayor precisión, utilizando técnicas estadísticas estándar sin necesidad de emplear modelos complejos.

1. PISA (Programme for International Student Assessment) es una evaluación internacional coordinada por la OCDE, que mide cada tres años el rendimiento de estudiantes de 15 años en lectura, matemáticas y ciencias, con el fin de evaluar la preparación para la vida adulta.
2. TIMSS (Trends in International Mathematics and Science Study) es un estudio internacional desarrollado por la IEA, que evalúa el rendimiento en matemáticas y ciencias de estudiantes de cuarto y octavo grado, permitiendo comparar los logros educativos.
3. ERCE es un estudio regional de aprendizaje realizado por el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE) de la UNESCO. Evalúa el desempeño de los estudiantes de primaria en América Latina y el Caribe, proporcionando información comparativa que apoya la formulación de políticas educativas.
4. Saber 3°, 5°, 7° y 9° es una evaluación estandarizada colombiana que mide el desarrollo de competencias básicas en estudiantes de educación primaria y secundaria, junto con factores asociados como contexto socioeconómico y habilidades socioemocionales.
5. Diseño en el cual a cada estudiante se le asigna una única forma o cuadernillo con un subconjunto de ítems del total disponible. Para más detalles, consultar: https://www.icfes.gov.co/wp-content/uploads/2025/02/2-Edicion-que-diseno-del-armado-se-emplea-en-el-icfes-para-medir-las-pruebas-saber-.pdf

Las características contextuales en este tipo de evaluaciones se recolectan a través de cuestionarios aplicados a los estudiantes y en algunos casos a docentes y directivos. Esta información permite comprender las condiciones educativas, sociales y personales que influyen en el aprendizaje de los estudiantes. Además del análisis descriptivo de estas características, estos datos se incorporan en el proceso de generación de valores plausibles, contribuyendo a mejorar la precisión en la estimación de la habilidad latente[6].

El proceso de generación de valores plausibles parte del principio de que no se puede estimar con precisión la habilidad individual de cada estudiante solo con sus respuestas a la prueba y recurre a un enfoque bayesiano en el cual la habilidad latente de cada estudiante se considera una variable aleatoria. Para estimar su distribución posterior se combinan dos fuentes de información: las respuestas a los ítems que presentó de la prueba y las características contextuales del estudiante *(como sexo, nivel educativo de los padres, acceso a recursos, etc.)* que se suelen recolectar en este tipo de estudios. El resultado es una distribución posterior de la habilidad para cada estudiante.

Esta distribución posterior tiene en cuenta tanto el modelo de medición *(basado en la Teoría de Respuesta al Ítem, TRI)* como un modelo de regresión que predice la habilidad esperada a partir de variables contextuales. El modelo de regresión se puede expresar como:

$$\theta_i = \mu_i + \varepsilon_i = X_i\Gamma + \varepsilon_i,$$

Donde: θ_i es la habilidad latente del estudiante i,

X_i es el vector de variables contextuales,
Γ es el vector de coeficientes de regresión y
$\varepsilon_i \sim \mathrm{N}(0, \Sigma)$ es un término de error aleatorio que se supone sigue una distribución normal.

Con base en este modelo, se estima para cada estudiante una distribución posterior condicional de su habilidad dada su cadena de respuestas *(string)* y contexto:

$$f(\theta_i | y_i X_i),$$

Donde: y_i son las respuestas del estudiante.

Luego, se toman múltiples muestras aleatorias desde esta distribución para cada estudiante. Cada muestra representa un posible valor de su habilidad y se denomina valor plausible. En las evaluaciones educativas internacionales, este proceso suele repetirse entre 5 y 10 veces, lo que permite generar una cantidad equivalente de valores plausibles para cada estudiante *(Bibby, 2020)*. Si bien teóricamente podrían generarse más valores, esto implicaría una mayor carga computacional sin necesariamente aportar mejoras significativas en la precisión de las estimaciones. Este conjunto de valores representa distintas posibles estimaciones de su habilidad, y refleja la incertidumbre inherente al proceso de medición individual.

La metodología en la práctica sigue estos pasos. Primero, se estiman los parámetros del modelo TRI para obtener la escala de habilidad, luego se ajusta un modelo de regresión de habilidades sobre los predictores contextuales X obteniendo Γ y Σ. A continuación, se calcula la distribución posterior $\mathrm{N}(\mu_{ip}, \Sigma_{ip})$ que depende de las respuestas y de la información de contexto. Por último, se extraen de esa distribución muestras aleatorias para obtener los valores plausibles:

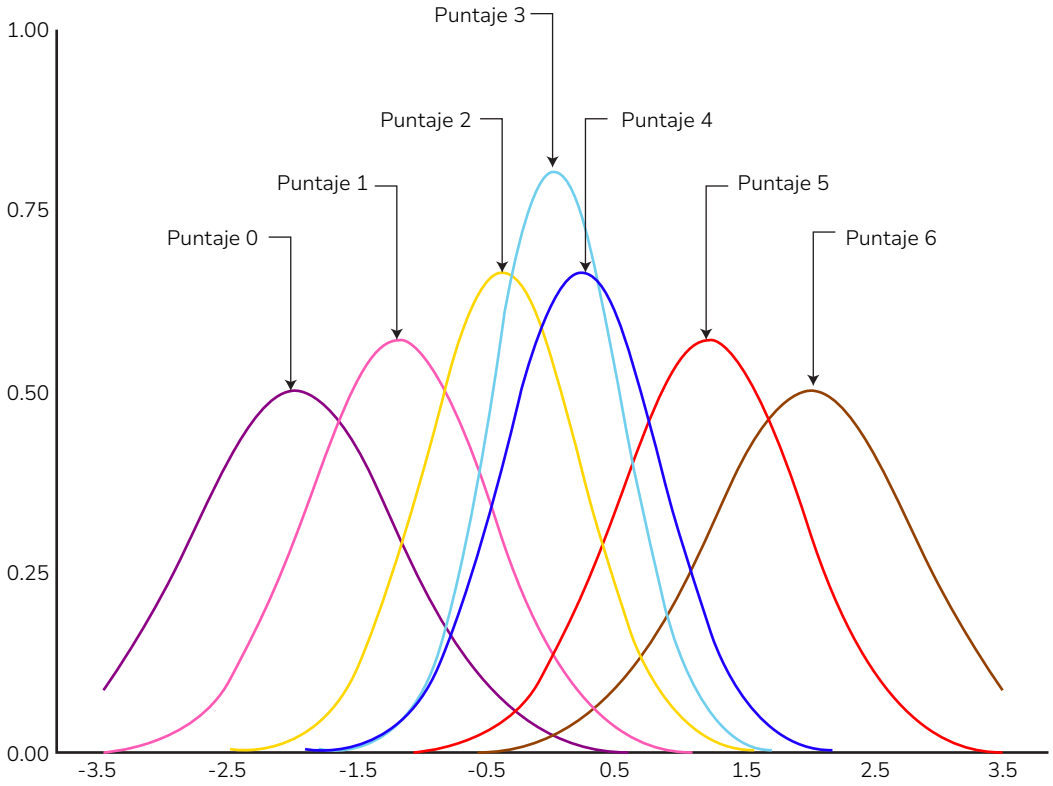
$$PV_{ij} \sim \mathrm{N}(\mu_{ip}, \Sigma_{ip}).$$

Cada uno de estos valores plausibles es tratado como una posible *"realización"* de la habilidad del estudiante, y se usan en análisis posteriores para reflejar la variabilidad e incertidumbre de las estimaciones a nivel poblacional.

6. Se entiende por habilidad latente una característica no observable directamente, pero que puede inferirse a partir del patrón de respuestas del estudiante en la prueba.

En la Figura 1 se muestra las distribuciones posteriores de habilidad para 7 diferentes puntuaciones obtenidas por los estudiantes en una evaluación. Se puede observar que, a medida que aumenta la puntuación del estudiante la distribución posterior se desplaza hacia la derecha, indicando un mayor nivel de habilidad estimada. Sin embargo, estas distribuciones se traslapan considerablemente, lo que evidencia que una misma puntuación puede estar asociada a diferentes niveles de habilidad con cierta probabilidad. Esto justifica la necesidad de generar múltiples valores plausibles a partir de estas distribuciones, en lugar de usar una única estimación puntual, ya que permite capturar la incertidumbre inherente en la medición de la competencia individual.

Figura 1
Habilidad en escala logit



Gráfica adaptada de Plausible Values por M. Wu, en Rasch Measurement Transactions, 18(2), 2004, pp. 976–978.

¿Cómo se hacen los cálculos con valores plausibles?

En las evaluaciones estandarizadas que utilizan valores plausibles, usualmente se generan P valores para cada estudiante en cada prueba o escala. Para calcular una estadística poblacional θ, por ejemplo, un promedio, un percentil o cualquier otra, la metodología indica que se debe realizar el cálculo por separado con cada uno de los P valores plausibles, y luego promediar los P resultados obtenidos.

Esto es,

$$\theta = \frac{1}{P}\sum_{j=1}^{P}\theta_j,$$

Donde: θ_j es la estadística poblacional de interés calculada con el j-ésimo valor plausible.

Adicionalmente, los valores plausibles permiten estimar la parte del error total atribuible a la incertidumbre en la imputación del rasgo latente θ, debido a la falta de información completa en la medición. Esta componente se conoce como varianza de imputación, y se calcula como la varianza entre las estimaciones obtenidas con cada valor plausible:

$$B_M = \frac{1}{P-1}\sum_{j=1}^{P}(\theta_j - \theta)^2.$$

Para ilustrar el uso de los valores plausibles, supongamos que tenemos una evaluación estandarizada a gran escala como Saber 3°, 5°, 7° y 9°, cuyo objetivo es estimar el desempeño de una subpoblación. En esta evaluación, se desea calcular el promedio de habilidades para una subpoblación con n estudiantes, donde estos solo responden una parte del total de ítems debido a que el contenido de la prueba es amplio y se empleó un diseño por bloques, lo que genera datos faltantes. Además, se utiliza información contextual obtenida por cuestionarios socioeconómicos para mejorar la estimación de la habilidad.

Aplicando modelos TRI y una regresión latente con estas variables, se generan los P valores plausibles por estudiante, que son muestras aleatorias de la distribución posterior de su habilidad. Inicialmente, calculamos el promedio para cada valor plausibles, así:

$$\bar{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} PV_{ij},$$

Donde: PV_{ij} es el valor plausible del i-ésimo estudiante de la población en el j-ésimo valor plausible con $j = 1, 2, \ldots, P$

Luego, se obtiene el promedio general combinando los promedios de los P valores plausibles:

$$\bar{\theta} = \frac{1}{P}\sum_{j=1}^{P}\bar{\theta}_j,$$

el cuál sería el puntaje promedio para la población de interés en nuestra evaluación. Por su parte, la varianza de imputación se calcularía como:

$$B_M = \frac{1}{P-1}\sum_{j=1}^{P}(\bar{\theta}_j - \bar{\theta})^2.$$

Siguiendo con el ejemplo[7], se desea estimar el promedio en el área de matemáticas para una subpoblación de 200 estudiantes de quinto grado, distribuidos en cuatro colegios. De acuerdo con la metodología adoptada, para este ejemplo se usan cinco valores plausibles (P=5) por estudiante.

7. Por razones prácticas, en el ejemplo se emplean de manera parcial algunos elementos de la prueba Saber 3°, 5°, 7° y 9°, como el grado, área y la escala. Sin embargo, las características y parámetros aplicados en el estudio completo se describen en una sección posterior, como la cantidad de valores plausibles y el uso de los pesos muestrales.

En la Tabla 1 se presentan los valores plausibles *(VP)* asignados a los estudiantes, junto con la sumatoria correspondiente para cada conjunto de valores y el cálculo del promedio asociado a cada uno.

Tabla 1
Ejemplo de valores plausibles para cálculo de la media.

Id_Colegio	Id_Estudiante	PV1	PV2	PV3	PV4	PV5
C1	1	498,2	442,4	490,6	445,9	466,2
C1	2	325,7	349,6	279,7	360,7	346,6
C1	3	407,1	457,0	444,7	422,0	489,1
C1	4	399,5	463,8	426,7	445,3	438,1
C1	5	374,9	439,0	406,3	437,8	433,2
C2	6	440,0	430,7	460,3	465,6	437,4
...	...	...	...	...	...	...
...	...	...	...	...	...	...
C10	195	388,7	416,9	418,5	406,6	388,4
C10	196	422,7	357,8	404,7	362,1	355,9
C10	197	338,0	364,1	342,9	350,8	375,0
C10	198	387,4	351,9	365,7	358,2	397,1
C10	199	402,2	358,5	380,3	395,4	374,1
C10	200	460,8	448,0	484,9	438,4	468,7
	∑PVij	78.361,3	78.731,9	78.508,9	79.150,4	78.342,1
	n	200	200	200	200	200
	Media	391,8	393,7	392,5	395,8	391,7

Luego la media para la subpoblación en matemáticas es 393,1 y se calcula así:

$$\bar{\theta} = \frac{1}{5}\sum_{j=1}^{5}\bar{\theta}_j = \frac{1}{5}\,(391{,}8 + 393{,}7 + \ 392{,}5 + \ 395{,}8 + \ 391{,}7) = \ 393{,}1$$

Por su parte, la varianza de imputación es:

$$B_M = \frac{1}{P-1}\sum_{j=1}^{P}\left(\bar{\theta}_j - \bar{\theta}\right)^2 = \frac{1}{5-1}((391{,}8-393{,}1)^2 + (393{,}7-393{,}1)^2 + (392{,}5-393{,}1)^2 + (395{,}8-393{,}1)^2 + (391{,}7-393{,}1)^2) = 2{,}915$$

Lo que no se debe hacer con los valores plausibles

Como se ha explicado en los apartados anteriores, la metodología de valores plausibles tiene un componente técnico importante, que requiere seguir los procedimientos para obtener resultados válidos. Un uso incorrecto de los valores plausibles puede llevar a conclusiones erróneas o sesgadas, por lo que es fundamental conocer qué prácticas deben evitarse *(von Davier, et. al, 2009).*

No se deben utilizar los valores plausibles para hacer reportes individuales, como asignar un puntaje único a un estudiante o comparar personas entre sí. Estas evaluaciones están diseñadas para hacer inferencias sobre poblaciones, no sobre individuos. Reportar un valor plausible como si fuera la *"habilidad verdadera"* de un estudiante puede ser impreciso e injusto, ya que los valores plausibles son el resultado de un proceso estadístico condicionado por respuestas y factores contextuales, y varían para reflejar la incertidumbre de esa estimación.

Tampoco se deben promediar los valores plausibles a nivel de estudiante para luego calcular una estadística con ese promedio. Aunque es una forma de trabajar con los datos, hacerlo elimina la variabilidad entre imputaciones y conduce a estimaciones sesgadas, especialmente en análisis con (TRI) logístico de dos parámetros (2PL)8, percentiles, correlaciones o regresiones. Ese promedio actúa como una estimación puntual del estudiante, lo cual contradice el propósito de los valores plausibles, que es representar la incertidumbre inherente en la medición.

¿Cómo aplicamos los valores plausibles en los exámenes desarrollados por el Icfes?

El Icfes implementa la metodología de valores plausibles en el estudio Saber 3°, 5°, 7° y 9°, una evaluación estandarizada a gran escala que se aplica a estudiantes de educación básica primaria y secundaria en Colombia. Su propósito es brindar información relevante sobre el avance del país en materia educativa e identificar el desarrollo de las competencias fundamentales y obligatorias de los estudiantes *(Icfes, 2023)*. En su versión actual, la evaluación no busca generar resultados individuales, sino ofrecer información confiable a nivel agregado, lo cual es valioso para la toma de decisiones en materia de políticas públicas educativas.

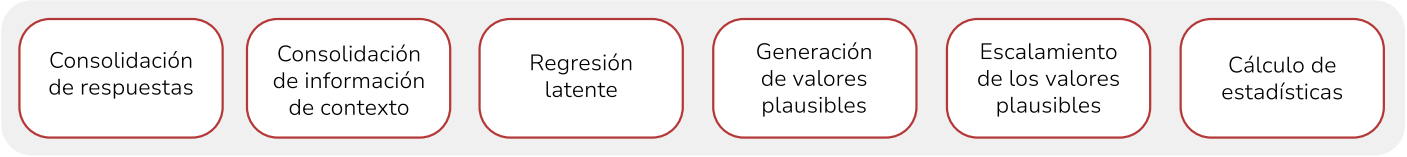
La población objetivo de esta evaluación está conformada por los estudiantes de 3°, 5°, 7° y 9° grado a nivel nacional, así como por ciertos grupos de interés, tales como el sector *(oficial y no oficial)*, la zona *(urbana y rural)* y otras subpoblaciones relevantes, como las regiones. Dado que realizar un censo resulta complejo, la evaluación se aplica actualmente a una muestra aleatoria de estudiantes que garantiza representatividad tanto a nivel nacional como en los principales agregados de interés. Por esta razón, se emplea un diseño muestral y factores de expansión para la generación de estimaciones.

La prueba se construye bajo un Diseño en Bloques Balanceado, por lo que cada estudiante responde solo a un subconjunto de los ítems del área evaluada. Junto con las pruebas cognitivas, se aplica un cuestionario que recoge información sobre el contexto socioeconómico del estudiante y otros factores asociados que pueden influir en su aprendizaje. La calificación de las pruebas cognitivas se realiza mediante un modelo de Teoría de Respuesta al Ítem (TRI) de dos parámetros (2PL)[8], donde cada ítem se caracteriza por dos propiedades: su dificultad, que indica el nivel de habilidad necesario para tener una alta probabilidad de responderlo correctamente, y su discriminación, que refleja qué tan bien el ítem diferencia entre estudiantes con distintos niveles de habilidad.

8. Para más detalles, consultar: https://www.icfes.gov.co/wp-content/uploads/2025/02/8-Edicion-boletin-saber-al-detalle.pdf

En la Figura 2 se presenta el esquema seguido en la prueba Saber 3°, 5°, 7° y 9°, para generar y usar los valores plausibles. Inicia con la consolidación de respuestas y de información contextual de los evaluados, sigue con la estimación de un modelo latente que relaciona desempeño y contexto, y continúa con la generación de 10 valores plausibles por estudiante. Luego, estos valores se transforman para estar en una misma escala y finalmente se usan para calcular estadísticas agregadas que informan la toma de decisiones educativas.

Figura 2
Esquema del proceso de estimación de valores plausibles en el examen Saber 3°, 5°, 7° y 9°.



Conclusiones

La metodología de valores plausibles es una herramienta estadística importante y necesaria en las evaluaciones estandarizadas a gran escala, ya que permite realizar inferencias válidas a nivel poblacional sin depender de estimaciones individuales que podrían estar sesgadas. La estrategia de generar múltiples valores de la distribución posterior de habilidad permite modelar la variabilidad del rasgo latente de manera más cercana a la realidad. Esta aproximación reduce el sesgo en la estimación de diferentes estadísticas *(medias, varianzas, percentiles y correlaciones)* y se ajusta a diseños de armado en bloques con datos faltantes.

En el Icfes la metodología de valores plausibles es una herramienta muy importante en el estudio Saber 3°, 5°, 7° y 9°, donde por el carácter diagnóstico y la extensión de contenido a evaluar, los estudiantes solo presentan un subconjunto de los ítems de las pruebas y se emplean valores plausibles para estimar la habilidad asociada a cada una de las áreas evaluadas en diferentes poblaciones. Este enfoque permite tener en cuenta la incertidumbre propia de medir habilidades que no se observan directamente, y así obtener resultados más precisos y representativos cuando se analizan grupos de estudiantes.

Referencias

- Bibby, Y. (2020). Plausible values: How many for plausible results? (Ph.D. thesis). University of Melbourne, Melbourne, Australia.
- Córdoba, M. (2016). Una aplicación de valores plausibles a la calificación de pruebas estandarizadas vía simulación. Comunicaciones en Estadística, 9(1), 55–78.
- Instituto Colombiano para la Evaluación de la Educación (Icfes). (2023). Informe descriptivo Saber 3°, 5°, 7° y 9° 2023. Icfes.
- Mullis, I. V. S., Martin, M. O., & Arora, A. (2025). TIMSS 2023 International results in mathematics and science. TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timss2023/international-results/
- Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2023). PIRLS 2021 International results in reading. TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/pirls2021/international-results/
- Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2023). PISA 2022 results (Volume I): The state of learning and equity in education. OECD Publishing. https://doi.org/10.1787/3d9c4dcd-en
- UNESCO. (2021). Estudio regional comparativo y explicativo (ERCE 2019). Oficina Regional de Educación para América Latina y el Caribe (OREALC/UNESCO Santiago).
- von Davier, M., González, E. J., & Mislevy, R. J. (2009). What are plausible values and why are they useful? En M. von Davier & D. Hastedt (Eds.), Issues and methodologies in large-scale assessments (Vol. 2, pp. 9–36). IERI Monograph Series.