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

Educación | Icfes

SABER al detalle

Edición 19

¿Cómo se procesan los resultados de evaluaciones a gran escala que usan valores plausibles y pesos de réplica?

Educación | Icfes

Saber Al Detalle:

¿Cómo se procesan los resultados de evaluaciones a gran escala que usan valores plausibles y pesos de réplica?

Subdirección de Estadísticas

Presidente de la República
Gustavo Francisco Petro Urrego

Ministro de Educación Nacional
José Daniel Rojas Medellín

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 Montaño Rincón

Subdirectora de Análisis y Divulgación
Alejandra Neira Aroca

Elaboración del documento
Juan Carlos Rubriche Cárdenas
Juan Felipe Salamanca Garzón

Revisado por
John Alexander Calderón Rodriguez

Diseño y diagramación
Linda Nathaly Sarmiento Olaya

Bogotá D.C., octubre de 2025

¿Cómo se procesan los resultados de evaluaciones a gran escala que usan valores plausibles y pesos de réplica?

Introducción

En las evaluaciones educativas a gran escala, como Saber 3°, 5°, 7° y 9° en Colombia y el Programa para la Evaluación Internacional de Estudiantes (PISA) de la OCDE, es fundamental utilizar valores plausibles para representar la incertidumbre inherente a la medición del desempeño cuando se infiere una habilidad no observada directamente (Wu, 2005) y pesos de réplica para estimar de manera robusta la variabilidad asociada a diseños muestrales complejos mediante métodos de replicación como BRR y su variante con ajuste de Fay (Judkins, 1990). Esto es coherente con el hecho de que este tipo de exámenes pretende medir características a nivel agregado (no individual) y con que los evaluados no presentan todos los módulos del instrumento; por tanto, estas técnicas permiten calcular estadísticas agregadas —como medias y proporciones— y sus errores estándar incorporando explícitamente tanto la incertidumbre de medición como la complejidad del muestreo (OCDE, 2009).

En lugar de asignar a cada estudiante un único puntaje "verdadero", se le asocia una distribución de probabilidad de habilidades. De esta distribución se extraen varios valores simulados (por ejemplo 10, para el caso de Saber 3°, 5°, 7° y 9°) llamados **valores plausibles (Icfes, 2025).** Estos valores se fundamentan en la teoría de imputación múltiple para inferencias en encuestas complejas (Mislevy, 1991). Cada valor plausible es una imputación del puntaje latente del estudiante. El uso de múltiples imputaciones de este tipo, siguiendo el marco teórico desarrollado por Rubin (1987), permite obtener estimaciones insesgadas de parámetros poblacionales y cuantificar la incertidumbre debida a la estimación de las habilidades. En palabras simples, los valores plausibles reflejan el hecho de que el desempeño real del estudiante es desconocido; por ello se generan varios puntajes posibles para cada evaluado, todos igualmente plausibles dado su patrón de respuestas.

Por otra parte, estas evaluaciones siguen un diseño muestral complejo, usualmente estratificado y por conglomerados (por ejemplo, primero se seleccionan colegios y luego estudiantes dentro de ellas). Como consecuencia, los estudiantes en la muestra no representan a igual cantidad de individuos de la población; cada uno tiene un peso de muestreo que indica a cuántos estudiantes representa. Para calcular errores estándar adecuados bajo estos diseños complejos, Saber 3°, 5°, 7° y 9° y PISA proporcionan conjuntos de pesos de réplica creados mediante el método de ***Réplicas Repetidas Balanceadas (BRR)*** (**Icfes, 2025).** Esta técnica permite la estimación robusta de varianzas en diseños muestrales estratificados sin asumir linealidad (Wolter,

2007). La idea del BRR es generar múltiples pseudomuestras alterando los pesos y recalculando la estadística de interés en cada réplica; la variabilidad entre estas réplicas sirve para estimar la varianza de muestreo. Cada réplica es un vector de pesos ajustados que repondera la misma muestra conforme al esquema BRR del diseño muestral original. Usando todas las réplicas, se calcula el error estándar de forma robusta sin requerir fórmulas analíticas complejas específicas del diseño. Para más detalles ver el Anexo sobre el diseño muestral.

Este documento, pretende hacer un breve repaso de los conceptos presentados en las dos anteriores ediciones de este boletín, donde se analiza en detalle los conceptos de Valores Plausibles (PV, por sus siglas en inglés), cuya metodología se describe en Saber al Detalle 17, y de Réplicas Repetidas Balanceadas (BRR, por sus siglas en inglés), desarrolladas en Saber al Detalle 18, así como la generación de estadísticas agregadas empleando estos conceptos y con ejemplos ilustrativos para que el lector pueda replicar de forma simple los resultados aquí expuestos.

¿Por qué son importantes los PV y BRR?

Al analizar las bases de datos de estas evaluaciones, ignorar los valores plausibles o los pesos replicados puede sesgar las estimaciones y producir errores estándar subestimados. Para obtener resultados válidos que reflejen la confiabilidad de las pruebas y la estructura de la muestra, se deben combinar tanto la variabilidad por muestreo como la variabilidad por imputación en los cálculos. Así, cualquier promedio, proporción u otra estadística debe calcularse de manera ponderada y replicada, y si involucra puntajes de desempeño, debe incorporar la repetición sobre los múltiples valores plausibles por estudiante.

Afortunadamente, existen herramientas prácticas en el software estadístico R que facilitan estos cálculos complejos. El ICFES ha desarrollado macros en R para automatizar el cálculo de estadísticos con pesos replicados y valores plausibles en sus bases de datos. De forma similar, en el contexto PISA, la comunidad dispone de paquetes en R como *intsvy* (International Survey Analysis) creados para manejar los datos de la OCDE y la IEA.

Cálculo de errores estándar

Tanto en Saber 3579 como en PISA, el cálculo completo del error estándar de una estimación se obtiene combinando dos componentes: la varianza entre réplicas (que refleja la incertidumbre muestral) y la varianza entre valores plausibles (que refleja la incertidumbre del puntaje).

En términos generales, sea $\hat{\theta}$ la estimación de interés (por ejemplo, una media poblacional, una proporción o un coeficiente de regresión) y supóngase que se dispone de P valores plausibles. Para cada valor plausible $i = 1, \dots,$ P se calcula una estimación puntual $\hat{\theta}_i$ usando el peso final. Adicionalmente, para cada valor plausible i y para cada peso de réplica $r = 1, \dots, R$ se obtiene una estimación replicada $\hat{\theta}_{i,r}$ utilizando el correspondiente peso replicado (esquema BRR).

Primero, para cada valor plausible i se estima la varianza de muestreo mediante el esquema de pesos de réplica. Para el valor plausible i, la varianza BRR se calcula como

$$V_{BRR}(\hat{\theta}_i) = \frac{1}{R(1-\mathrm{k})^2} \sum_{r=1}^{R} (\hat{\theta}_{i,r} - \hat{\theta}_i)^2,$$

donde R es el número de réplicas y k es el factor de Fay utilizado en la construcción de los pesos de réplica ($k = 0$ es un BRR balanceado estándar y, por ejemplo, $k = 0{,}5$ es un diseño Fay-BRR). Estas P varianzas de réplica capturan únicamente la incertidumbre debida al muestreo complejo.

Luego, se combinan los resultados de los P valores plausibles. Sea

$$\bar{\theta} = \frac{1}{P}\sum_{i=1}^{P}\hat{\theta}_i$$

la estimación del parámetro a nivel poblacional y

$$V(\hat{\theta}_1, \dots, \hat{\theta}_P) = \frac{1}{P-1}\sum_{i=1}^{P}(\hat{\theta}_i - \bar{\theta})^2$$

la varianza muestral de los P resultados puntuales obtenidos con cada valor plausible (varianza entre valores plausibles). Esta componente mide la incertidumbre asociada a la imputación de los puntajes.

La varianza total de la estimación combinada se obtiene entonces como

$$V(\bar{\theta}) = \frac{1}{P}\sum_{i=1}^{P}V_{BRR}(\hat{\theta}_i) + \left(1 + \frac{1}{P}\right)V(\hat{\theta}_1, \dots, \hat{\theta}_P)$$

El factor $(1 + {}^{1}/_{P})$ es una corrección por el uso de un número finito de imputaciones. De esta forma, los errores estándar combinados reflejan tanto la dispersión debido al muestreo como la debida a la estimación de los puntajes de los evaluados.

El siguiente ejemplo pequeño ilustra el procedimiento para calcular el error estándar de la media estimada $\hat{\mu}$ de una prueba cognitiva. Supongamos que hay 5 estudiantes y 3 valores plausibles:

Est. j	Peso final W_j	PV1 μ_{j1}	PV2 μ_{j2}	PV3 μ_{j3}
1	8	40	38	42
2	12	45	48	46
3	10	50	52	48
4	15	60	58	57
5	5	55	60	53

Además, se asume que se produjeron con la metodología BRR los 4 conjuntos de pesos de réplica (los mismos para todos los PV):

Est. j	W_j	R1 $W_j^{(1)}$	R2 $W_j^{(2)}$	R3 $W_j^{(3)}$	R4 $W_j^{(4)}$
1	8	4	4	12	12
2	12	6	6	18	18
3	10	5	5	15	15
4	15	22,5	22,5	7,5	7,5
5	5	2,5	7,5	2,5	7,5

Usando los valores plausibles se calculan las estimaciones de las medias $\hat{\mu}_p$ para cada estudiante con $p = 1, \ldots, P$ y se obtiene la media combinada, así:

$$\hat{\mu}_{p,r} = \frac{\sum_j w_j^{(r)} \mu_{jp}}{\sum_j w_j^{(r)}}$$

- $\hat{\mu}_1 = \frac{8*40+12*45+10*50+15*60+5*55}{50} = 50{,}70$
- $\hat{\mu}_2 = \frac{8*38+12*48+10*52+15*58+5*60}{50} = 51{,}40$
- $\hat{\mu}_3 = \frac{8*42+12*46+10*48+15*57+5*53}{50} = 49{,}76$

con estimador de la media poblacional de los puntajes:

$$\bar{\mu} = \frac{\hat{\mu}_1 + \hat{\mu}_2 + \hat{\mu}_3}{3} = \frac{50{,}70+51{,}40+49{,}76}{3} = 50{,}62$$

Entonces, la varianza entre valores plausibles es:

$$V(\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = \frac{1}{3-1} \sum_{i=1}^{m} (\hat{\mu}_i - \bar{\mu})^2 = 0{,}6772$$

Ahora, se calculan las medias replicadas usando:

$$\hat{\mu}_{p,r} = \frac{\sum_j w_j^{(r)} \mu_{jp}}{\sum_j w_j^{(r)}}$$

» Para el primero conjunto de pesos de réplica $r = 1$.

$$\hat{\mu}_{1,1} = \frac{4*40+6*45+5*50+22{,}5*60+2{,}5*55}{40} = 54{,}19$$

$$\hat{\mu}_{2,1} = \frac{4*38+6*48+5*52+22{,}5*58+2{,}5*60}{40} = 53{,}88$$

$$\hat{\mu}_{3,1} = \frac{4*42+6*46+5*48+22{,}5*57+2{,}5*53}{40} = 52{,}48$$

» Para el segundo conjunto de pesos de réplica $r = 2$.

$$\hat{\mu}_{1,2} = \frac{4*40+6*45+5*50+22{,}5*60+7{,}5*55}{45} = 54{,}28$$

$$\hat{\mu}_{2,2} = \frac{4*38+6*48+5*52+22{,}5*58+7{,}5*60}{45} = 54{,}56$$

$$\hat{\mu}_{3,2} = \frac{4*42+6*46+5*48+22{,}5*57+7{,}5*53}{45} = 52{,}53$$

» Para el tercer conjunto de pesos de réplica $r = 3$.

$$\hat{\mu}_{1,3} = \frac{12*40+18*45+15*50+7{,}5*60+2{,}5*55}{55} = 47{,}77$$

$$\hat{\mu}_{2,3} = \frac{12*38+18*48+15*52+7{,}5*58+2{,}5*60}{55} = 48{,}82$$

$$\hat{\mu}_{3,3} = \frac{12*42+18*46+15*48+7{,}5*57+2{,}5*53}{55} = 47{,}49$$

» Para el cuarto conjunto de pesos de réplica $r = 4$.

$$\hat{\mu}_{1,4} = \frac{12*40+18*45+15*50+7{,}5*60+7{,}5*55}{60} = 48{,}38$$

$$\hat{\mu}_{2,4} = \frac{12*38+18*48+15*52+7{,}5*58+7{,}5*60}{60} = 49{,}75$$

$$\hat{\mu}_{3,4} = \frac{12*42+18*46+15*48+7{,}5*57+7{,}5*53}{60} = 47{,}95$$

Ahora, se hayan las varianzas BRR de las estimaciones replicadas con $R = 4$ y con constante de Fay $k = 0{,}5$ mediante:

$$V_{BRR}(\hat{\mu}_i) = \frac{1}{4(1-0{,}5)^2} \sum_{r=1}^{R} (\hat{\mu}_{i,r} - \hat{\mu}_i)^2 = \sum_{r=1}^{4} (\hat{\mu}_{i,r} - \hat{\mu}_i)^2$$

Por lo tanto,

- $V_{BRR}(\hat{\mu}_1) = (54{,}19 - 50{,}70)^2 + (54{,}28 - 50{,}70)^2 + (47{,}77 - 50{,}70)^2 + (48{,}38 - 50{,}70)^2 = 38{,}96$
- $V_{BRR}(\hat{\mu}_2) = (53{,}88 - 51{,}40)^2 + (54{,}56 - 51{,}40)^2 + (48{,}82 - 51{,}40)^2 + (49{,}75 - 51{,}40)^2 = 25{,}51$
- $V_{BRR}(\hat{\mu}_3) = (52{,}48 - 49{,}76)^2 + (52{,}53 - 49{,}76)^2 + (47{,}49 - 49{,}76)^2 + (49{,}95 - 49{,}76)^2 = 20{,}26$

Finalmente, la varianza total de la estimación combinada es:

$$V(\bar{\mu}) = \frac{1}{3}\sum_{i=1}^{3} V_{BRR}(\hat{\mu}_i) + \left(1 + \frac{1}{3}\right) V(\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = \frac{1}{3}(38{,}96 + 25{,}51 + 20{,}26) + \frac{4}{3}(0{,}6772) = 29{,}15$$

Por lo tanto, el error de estimación de la media, calculado como la raíz cuadrada de su varianza estimada, es $\sqrt{V(\bar{\mu})} = 5{,}40$.

Procesamiento de resultados en software estadístico R

A continuación, se presentan orientaciones para procesar los datos de Saber 3579 y de PISA en R, utilizando las herramientas disponibles para cada caso. Se asume que el lector ya descargó las bases de datos correspondientes: por ejemplo, un archivo .Rdata o .sav con la base de Saber 3579, o los archivos de datos de PISA (que pueden estar en formatos como .csv, .sav o .txt conforme los publica la OCDE).

Análisis de Saber 3579 con macros del ICFES

El ICFES provee un conjunto de funciones (a veces llamadas macros) en R para facilitar el cálculo de estadísticas con los datos de Saber 3579, estas se pueden encontrar en la plataforma *DataIcfes*. Para acceder a la misma, el usuario debe llenar un formulario proporcionado por el ICFES, relacionar un correo, y al mismo, llegará un mensaje, en donde se proporcionan los enlaces necesarios.

Estas funciones automatizan la lectura de los 10 valores plausibles y 84 réplicas de peso, devolviendo directamente las estimaciones con sus errores estándar. El archivo que contiene las funciones se denomina agreFunctions2.R y puede encontrarse en los scripts provistos por ICFES. Antes de usar las macros, asegúrese de instalar y cargar los paquetes R requeridos (Por ejemplo: dplyr, purrr, broom, tidyr, sjmisc, fastDummies, rlang).

Pasos para usar las macros del ICFES

1. Cargar la base de datos de Saber 3579: el ICFES proporciona la base consolidada como un archivo R, por ejemplo DATA_MEDICION_SB3579.Rdata, se puede cargar así:

Script de R

```
load("ruta/al/archivo/DATA_MEDICION_SB3579.Rdata")
# Supongamos que el objeto cargado es un data.frame llamado 'saber'.
```

La base saber debe contener, entre otras, las columnas de peso y réplicas (PESO, Peso_R1...Peso_R84), las columnas de valores plausibles (PV1, ... , PV10 por cada área) y las variables demográficas o de desagregación (grado, sexo, región, etc.). Verifique esto con names(saber) o str(saber).

2. Importar las funciones/macros de cálculo: Ubique el archivo agreFunctions2.R provisto por ICFES (macros de Saber 3579 del portal DataIcfes). Cárguelo en R con source() para tener disponibles las funciones. Por ejemplo:

Script de R

```
source("ruta/a/src/agreFunctions2.R")
# Esto define funciones como brr_mean_PV, brr_PORC_NoPV, entre otras.
```

Nota: Estas funciones se dividen en versiones con PV (que utilizan valores plausibles) y con NoPV (para variables sin valores plausibles). Por ejemplo, la función brr_mean_PV calcula la media de una variable de desempeño (usando todos los PV), mientras que la función brr_mean_NoPV calcula la media de una variable observada común. De igual forma, brr_PORC_NoPV obtiene proporciones (porcentajes) de categorías en una variable categórica.

Actualmente existen funciones para medias, desviaciones estándar, percentiles, regresiones, correlaciones, diferencias de medias y proporciones, tanto con valores plausibles como sin valores plausibles.

3. Calcular una media con valores plausibles: Suponga que se desea estimar la puntuación promedio en Lectura por sexo en la prueba Saber 3579. En la base, las columnas de valores plausibles de Lectura podrían llamarse LectPV1, ... , LectPV10 (según la notación vista en la documentación). La función apropiada es brr_mean_PV. Sus argumentos principales son: la base de datos de entrada, el nombre del peso muestral, una expresión regular para identificar los pesos replicados, otra para los PV, y opcionalmente la(s) variable(s) de desagregación (byvar). Usamos byvar="SEXO" asumiendo que existe una variable de sexo codificada (por ejemplo, SEXO con valores F/M). El código sería:

Script de R

```
# Media de Lectura por sexo utilizando 10 PV y BRR
resultado_media_lectura <- brr_mean_PV(
    infile = DATA_MEDICION_SB3579,
    weight = "PESO", # nombre de la columna de peso final
    repli_root = "Peso_R", # patrón para columnas de réplica
    pv_root = "LectPV", # patrón para valores plausibles de Lectura
    byvar = c("GRADO","SEXO"), # desagregar por grado y sexo (opcional; puede ser vector de
    varias vars)
    siNA = FALSE # ignorar NA's en sexo, si queremos excluir faltantes
    )
print(resultado_media_lectura)
```

Esta función internamente recorrerá los 10 valores plausibles de Lectura y los 84 pesos replicados para computar la media ponderada de Lectura por sexo y grado y su error estándar combinado. En la salida, típicamente obtendremos un data frame donde cada fila corresponde a una categoría de sexo (por ejemplo, Femenino, Masculino), a una categoría de grado (por ejemplo, 3, 5, 7 y 9), la media estimada de Lectura y la desviación estándar de esa estimación, para cada caso. Por ejemplo:

Grado	Sexo	mediaPV	media_SEFinal_PV	PVj
3	Femenino	390,103176	2,050259201	LectPV
3	Masculino	385,912479	2,225962691	LectPV
5	Femenino	391,972648	2,184051944	LectPV
5	Masculino	384,823331	2,868361543	LectPV
7	Femenino	404,779542	2,305652889	LectPV
7	Masculino	395,462355	2,128157465	LectPV
9	Femenino	371,319934	2,43388883	LectPV
9	Masculino	367,173551	1,841817866	LectPV

La columna mediaPV representaría la puntuación media estimada, calculada correctamente considerando los pesos, y media_SEFinal_PV sería el error estándar que ya incorpora la variabilidad entre réplicas y entre valores plausibles.

Internamente la función brr_mean_PV calcula la media con el peso final PESO para cada valor plausible de Lectura, luego repite el cálculo con cada réplica de peso Peso_Rk () para estimar la varianza de muestreo. Finalmente combina ambas fuentes de varianza para producir un error estándar total. Así, el resultado es comparable a lo que se obtendría con software especializado o macros oficiales.

4. Calcular una proporción: Ahora suponga que se quiere estimar la proporción de estudiantes por nivel socioeconómico (NSE) en la muestra de Saber 3579, a nivel nacional. La variable NSE en la base indica el nivel socioeconómico (Bajo, Medio, Alto). Para calcular proporciones de una variable categórica, usamos la función brr_PORC_NoPV (no hay PV involucrados, pues NSE no es un puntaje imputado sino un dato observado). Podemos dejar byvar = NULL para obtener el total nacional sin desagregar por otra variable. Ejemplo:

Script de R

```
# Distribución porcentual por nivel socioeconómico (NSE) a nivel nacional
resultado_prop_NSE <- brr_PORC_NoPV(
    infile = DATA_MEDICION_SB3579,
    weight = "PESO",
    repli_root = "Peso_R",
    varDom = "NSE", # variable de dominio para proporciones
    byvar = NULL, # NULL = cálculo a nivel total, sin desagregación adicional
    siNA = FALSE # ignorar valores perdidos de NSE, si los hubiera
    )
print(resultado_prop_NSE)
```

La salida de brr_PORC_NoPV presentará cada categoría de la variable NSE con su porcentaje estimado y error estándar. Por ejemplo:

Categoria	Porcentaje	porcentaje_SEM
NS1	0,254410097	0,008628929
NS2	0,393571277	0,007257212
NS3	0,249548361	0,008757885
NS4	0,063888404	0,0039305

Esto indicaría, por ejemplo, que un 25.4% de los estudiantes evaluados provienen del nivel socioeconómico “NS1”, con un error estándar de 0.86 puntos porcentuales, etc. Nuevamente, estos errores estándar reflejan la incertidumbre debida al muestreo (a través de los pesos replicados). Si quisiéramos desagregar la distribución por otra variable (por ejemplo, por grado o por región), podríamos añadir byvar = “GRADO” o byvar = c(“REGION”,”SEXO”) según la necesidad, y la función generaría una tabla de doble entrada con porcentajes por grupo.

5. Verificar y manejar datos faltantes: Las macros del ICFES tienen el parámetro siNA para controlar cómo tratar los valores faltantes (NA). Por defecto es TRUE (lo que significa que incluye una categoría explícita para NA en la salida). Sin embargo, suele ser recomendable poner siNA = FALSE para excluir los casos con datos faltantes de las variables de agrupación, evitando interpretaciones erróneas. En el caso de variables continuas (p.ej. calcular una media), las funciones internamente ya omiten los NA en la variable de interés.

Con estos pasos, el lector puede obtener de manera práctica y rápida estadísticas descriptivas de Saber 3579. Las macros cubren desde estadísticas simples (medias, porcentajes) hasta análisis más complejos (diferencias de medias/proporciones, regresiones con PV, correlaciones), todos

incorporando correctamente los 10 valores plausibles y el esquema de réplicas BRR. Esto garantiza que los resultados presentados tengan intervalos de confianza y pruebas de significancia correctamente calculados bajo el diseño muestral complejo de la evaluación.

Conclusiones

En esta entrega de **Saber AI Detalle** se ha presentado una guía práctica para procesar resultados de evaluaciones educativas de gran escala (Saber 3°,5°,7°,9°) usando R, incorporando valores plausibles y pesos replicados bajo diseños de muestra complejos. El uso correcto de valores plausibles y réplicas balanceadas permite a los lectores aprovechar al máximo la riqueza de las bases de datos de evaluaciones como Saber 3579 y PISA, obteniendo conclusiones válidas sobre el desempeño de los evaluados. Con las herramientas en R descritas, el proceso se vuelve reproducible y transparente, facilitando análisis secundarios, comparaciones internacionales o seguimiento de tendencias con rigor estadístico.

Anexos

Diseño muestral de Saber 3°, 5°, 7° y 9°

Saber 3°, 5°, 7° y 9° es una evaluación nacional colombiana que sigue un diseño de muestra probabilístico estratificado y por conglomerados en múltiples etapas. De manera simplificada, el proceso fue:

- **Estratificación y etapas:** Se estratificaron las escuelas (por ejemplo, por región geográfica, zona rural/urbana, sector público/privado, etc.) y luego se seleccionaron aleatoriamente establecimientos educativos. Dentro de cada establecimiento, posiblemente se seleccionaron sedes (muchas instituciones tienen varias sedes o campus) y finalmente estudiantes dentro de cada sede. Así, podemos pensar en hasta tres niveles: establecimiento, sede y estudiante. Este esquema asegura representatividad a nivel nacional y por subgrupos de interés, pero introduce complejidad en el cálculo de varianzas.

- **Pesos de muestreo:** Dado que no todos los estudiantes tienen la misma probabilidad de ser elegidos, a cada estudiante se le asigna un peso de expansión. El peso inicial se basó en las inversas de las probabilidades de selección en cada etapa (establecimiento, sede, estudiante). Luego el ICFES aplicó ajustes de calibración usando datos poblacionales (p.ej. el SIMAT, que es el sistema de matrículas) y correcciones por no respuesta, resultando en un peso final calibrado denominado típicamente PESO. Este peso permite que la muestra represente correctamente a la población objetivo de estudiantes en los grados evaluados.

- **Réplicas de peso (BRR):** Para calcular errores estándar bajo este diseño, el ICFES generó 84 pesos replicados (Peso_R1 a Peso_R84) para cada estudiante. Estas réplicas siguen el método BRR con el ajuste de Fay (un modificador que mejora la estabilidad de ciertas estimaciones como cuantiles). En términos prácticos, cada réplica es un conjunto alternativo de pesos ligeramente modificados (algunos estudiantes tienen su peso aumentado y otros reducido en cada réplica, según un esquema balanceado por estratos). Para cada réplica r se pueden recalcular las estadísticas de interés; la variación entre las estimaciones de las 84 réplicas permite estimar la varianza de muestreo. El uso de 84 réplicas en Saber 3579 obedece al número de pares de unidades primarias (escuelas o sedes) por estrato en el diseño de esa evaluación.

- **Valores plausibles:** Saber 3579 evaluó varias áreas (Matemáticas, Lectura, Ciencias Naturales y Competencias Ciudadana). Para cada área, en lugar de un puntaje observado único, el ICFES reportó 10 valores plausibles por estudiante. Estos PV se generaron usando modelos de Teoría de Respuesta al Ítem (modelo logístico de

2 parámetros, 2PL) calibrados con los datos de respuestas de los estudiantes. Por ejemplo, las variables MatePV1, … , MatePV10 corresponden a los diez valores plausibles de Matemáticas, LectPV1, … , LectPV10 a Lectura, y así sucesivamente.

De acuerdo con lo anterior, las bases de datos de resultados de Saber 3579 contienen columnas de identificación, variables de estratificación (ej. región, departamento, zona, sector), un peso final PESO, 84 pesos replicados Peso_R1, … , Peso_R84 y 40 valores plausibles en total (10 por cada una de las 4 áreas evaluadas) junto con otras respuestas de contexto. Con esta estructura, un investigador puede obtener estimaciones nacionales o por subgrupos calculando estadísticas con el peso PESO, y obteniendo errores estándar combinando los resultados de las réplicas y de los distintos PV.

Diseño muestral de PISA

La evaluación PISA tiene un diseño complejo conceptualmente similar, aunque con sus propias especificidades internacionales:

- **Estratificación y dos etapas:** En PISA, cada país realiza un muestreo estratificado de centros educativos (escuelas) primero, y luego dentro de cada escuela seleccionada, se elige aleatoriamente a estudiantes de 15 años para participar. Las estratificaciones suelen considerar regiones del país, tipo de escuela (pública/privada) y otras variables para asegurar representatividad. El diseño estándar de PISA es de dos etapas (escuela, estudiante) y altamente estratificado; no se utilizan normalmente más de dos etapas porque se parte de marcos de escuelas nacionales. Sin embargo, debido a las diferencias en tamaños de escuelas, a veces se aplica muestreo sistemático dentro de escuela o ponderación por tamaño para igualar cargas muestrales.

- **Pesos:** Cada estudiante en PISA tiene un peso de muestreo final llamado típicamente W_FSTUWT (Weight Final Student) que combina la probabilidad de selección de su escuela y de él dentro de la escuela, con ajustes postestratificación y por no respuesta. Estos pesos calibran la muestra de PISA para que reproduzca la población de 15 años de cada país.

- **Pesos replicados (BRR):** PISA provee 80 pesos de réplica por estudiante (W_FSTR1 a W_FSTR80 en los microdatos, a veces denominados final student replicate weights). Estos 80 pesos de réplica se construyen mediante el método BRR con factor de Fay (generalmente con coeficiente de Fay = 0.5) aplicado a las escuelas como unidades primarias. En PISA 2022, por ejemplo, se incluyeron 10 valores plausibles por área y 80 réplicas de peso en la base de datos. El procedimiento para usar estas réplicas es análogo al de Saber: se recalcula la estadística de interés con cada una de las 80 variantes de pesos y se combina la variabilidad de esos cálculos para obtener la varianza debida al muestreo.

- **Valores plausibles:** PISA evalúa tres áreas principales (Lectura, Matemáticas, Ciencias) y a veces áreas innovadoras dependiendo del año

(por ejemplo, Competencia Global, Resolución Colaborativa de Problemas, etc.). Para cada área principal, al igual que Saber, PISA asigna múltiples valores plausibles por alumno (tradicionalmente 5 PV por área en ediciones antiguas, extendidos a 10 PV por área en ediciones recientes para mayor precisión). Estos valores plausibles se nombran en la base con prefijos como PV1MATH...PV10MATH (Matemáticas), PV1READ... (Lectura) y PV1SCIE... (Ciencias), por ejemplo. Los PV se generan usando un modelo estadístico basado en la TRI y respuestas del estudiante a ítems cognitivos, combinado con información de contexto, para lograr imputaciones adecuadas de la habilidad latente. Igual que en Saber, los PV permiten incorporar la incertidumbre de estimar el puntaje verdadero del alumno.

Como resultado, PISA nos entrega bases de datos internacionales (por país o concatenadas) que incluyen: identificadores, variables de contexto (género, ESCS índice socioeconómico, etc.), peso final de estudiante W_FSTUWT, 80 pesos replicados W_FSTR1-W_FSTR80 y valores plausibles de desempeño en las áreas evaluadas. Cualquier análisis (por ejemplo, calcular la media de Matemáticas de un país, o la proporción de estudiantes que alcanzan cierto nivel) debe usar W_FSTUWT como peso principal y las 80 réplicas para errores estándar, y en el caso de puntajes de desempeño, iterar sobre todos los PV para combinar la incertidumbre de medición.

Referencias

Icfes. (2025). *Saber Al Detalle (edición 17): ¿Qué son los valores plausibles y cómo se implementan en el Icfes?* Bogotá D.C.

Icfes. (2025). *Saber Al Detalle (edición 18): Replicas repetidas balanceadas (BRR).* Bogotá D.C.

Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6(3), 223–239.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2), 177–196.

OCDE. (2009). *PISA Data Analysis Manual: SAS (2nd ed.).* OECD Publishing.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Wolter, K. M. (2007). *Introduction to Variance Estimation ((2nd ed.)*. Springer.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, 31(2–3), 114–128.