REPORTE TÉCNICO 3

ESTIMACIÓN DE VARIANZA EN ARISTAS

INEEd
Instituto Nacional de Evaluación Educativa

La elaboración de este documento estuvo a cargo de: Betania Ávalos, Meliza González y Juan José Goyeneche

Corrección de estilo: Mercedes Pérez y Federico Bentancor
Diseño y diagramación: Diego Porcelli

Montevideo, 2025

Edificio Los Naranjos, Planta Alta, Parque de Innovación del LATU
Av. Italia 6201, Montevideo, Uruguay
(+598) 2604 4649 – 2604 8590
ineed@ineed.edu.uy
www.ineed.edu.uy

Cómo citar: INEEd. (2025). *Reporte técnico 3. Estimación de varianza en Aristas*. https://www.ineed.edu.uy/images/publicaciones/reportes/Reporte-tecnico-3-Estimacion-varianza-Aristas.pdf

Este informe trata de adolescentes y adultos mujeres y varones. El uso del masculino genérico obedece a un criterio de economía de lenguaje y procura una lectura más fluida, sin ninguna connotación discriminatoria.

ÍNDICE

INTRODUCCIÓN

La prueba Aristas, tal como otras evaluaciones educativas de gran escala, se basa en un diseño muestral complejo, que puede describirse como estratificado bietápico por conglomerados. En la primera etapa se seleccionan centros educativos estratificados según características del centro (como categoría y región geográfica), mientras que en la segunda etapa se eligen grupos de estudiantes al interior de los centros seleccionados. Durante el levantamiento de la información, las unidades de muestreo están sujetas a no respuesta en ambas etapas. Posteriormente, los pesos finales consideran tanto los pesos de base, derivados de las probabilidades de selección, como los ajustes por no respuesta, entre otros aspectos.

La combinación de un diseño muestral complejo y los ajustes de ponderación por no respuesta, junto con posibles procedimientos de postestratificación, así como el tipo de estimador a calcular (por ejemplo, promedios, cuantiles, coeficientes de regresión), impacta directamente en el nivel de error de las estimaciones estadísticas (Atasever et al., 2025; Valliant et al., 2018).

El error muestral de estimación o error estándar, definido como la raíz cuadrada de la varianza muestral, cuantifica la precisión con que los parámetros son estimados y es central para la inferencia estadística. En este contexto, es necesario contar con una estimación precisa de este error para realizar estimaciones estadísticas confiables. Esto es indispensable para la correcta aplicación de pruebas de significación estadística y la construcción de intervalos de confianza. A su vez, coloca en un lugar central la elección de la metodología utilizada para estimar la varianza en el marco de un diseño complejo, como el que plantean las evaluaciones educativas a gran escala.

Existen distintas metodologías para estimar la varianza muestral, de las cuales tres de las principales son las siguientes: fórmulas exactas, linealización y técnicas de replicación (Valliant et al., 2018). En el marco de diseños muestrales complejos se ha popularizado la utilización de esta última (Valliant y Dever, 2018). Entre las evaluaciones educativas internacionales de gran escala, las principales metodologías de replicación utilizadas son: réplicas repetidas balanceadas (balanced repeated replication, BRR) con variante Fay, usada en PISA, TALIS (Atasever et al., 2025) y ERCE[1] (UNESCO, 2015), entre otras, y el método jackknife (JK2), usado en TIMSS, PIRLS, ICCS e ICILS[2] (Atasever et al., 2025).

[1] El Programa para la Evaluación Internacional de Alumnos (PISA, por su sigla en inglés) y el Estudio Internacional de Enseñanza y Aprendizaje (TALIS, por su sigla en inglés) son evaluaciones realizadas por la Organización para la Cooperación y el Desarrollo Económicos (OCDE). ERCE es el Estudio Regional Comparativo y Explicativo, organizado por el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE) bajo la coordinación de la Oficina Regional de Educación de UNESCO para América Latina y el Caribe. Para mayor información, consultar aquí: https://www.oecd.org/en/about/programmes/pisa.html, https://www.oecd.org/en/about/programmes/talis.html y https://www.unesco.org/es/fieldoffice/santiago/expertise/llece.

[2] El Estudio Internacional de Tendencias en Matemáticas y Ciencias (TIMSS, por su sigla en inglés), el Estudio Internacional de Progreso en Comprensión Lectora (PIRLS, por su sigla en inglés), el Estudio Internacional sobre Educación Cívica y Ciudadana (ICCS, por su sigla en inglés) y el Estudio Internacional sobre Alfabetización en Computación e Información (ICILS, por su sigla en inglés) son distintas evaluaciones internacionales realizadas por la Asociación Internacional para la Evaluación de Rendimiento Educativo (IEA, por su sigla en inglés). Para mayor información, consultar aquí: https://www.iea.nl/.

En consonancia con las prácticas realizadas en evaluaciones internacionales con diseños muestrales similares, las aplicaciones de Aristas han utilizado hasta el momento la metodología BRR-Fay. Debido a que este método no permite incorporar correcciones por población finita y dado que, en el contexto uruguayo, la población objetivo de Aristas no es de gran tamaño y las fracciones de muestreo son no despreciables e incluso relativamente altas en algunos estratos, resulta pertinente abordar este tema.

Como parte del monitoreo y la revisión permanente de los procedimientos que preservan la rigurosidad de Aristas, el presente reporte busca realizar una evaluación del desempeño de la metodología BRR-Fay. Al mismo tiempo, apunta a explorar alternativas metodológicas que permitan incluir explícitamente la corrección por fracción de muestreo utilizada en cada estrato.

Para lograr el objetivo propuesto, se compararon los resultados de tres metodologías de estimación de varianza: 1) replicación por el método BRR-Fay, 2) fórmulas exactas basadas en el diseño muestral sin uso del factor de corrección para poblaciones finitas y 3) fórmulas exactas basada en el diseño con uso del factor de corrección para poblaciones finitas. A su vez, en este último caso se exploraron variantes del factor de corrección.

Para evaluar el desempeño de estos métodos, este trabajo tomó como referencia varios aportes del estudio de la Asociación Internacional para la Evaluación de Rendimiento Educativo (IEA, por su sigla en inglés) desarrollado por Atasever et al. (2025), donde comparan el funcionamiento de metodologías de estimación de varianza en evaluaciones internacionales de gran escala. A diferencia de dicho trabajo, el presente análisis se centra en una población de menor tamaño (≈50.000 estudiantes frente a ≈500.000), lo que hace que la corrección por población finita adquiera relevancia, introduciendo particularidades metodológicas en los diseños evaluados y en la interpretación de los resultados.

En la sección de discusión se ofrecen recomendaciones prácticas derivadas de los resultados con los aportes del estudio y sus limitaciones.

MÉTODO

Para llevar a cabo el estudio, se simuló una población única que reproduce características de centros, grupos y estudiantes del marco muestral de Aristas Media 2025. A partir de esta población simulada, se extrajeron múltiples muestras representativas mediante un procedimiento de simulación Monte Carlo, sobre las cuales se aplicaron distintas metodologías de estimación de varianza para evaluar su comportamiento. Los resultados se analizaron mediante cálculos de diferentes estimadores en una selección de variables de interés relacionadas con el desempeño en las pruebas de lectura y matemática y con puntajes en índices sobre habilidades socioemocionales. Todos los análisis se realizaron con el software R, versión 4.4.2.

SIMULACIÓN DE DATOS DE LA POBLACIÓN

Para el marco muestral de Aristas Media 2025, compuesto por 586 centros educativos y 1.953 grupos de noveno grado, se simuló la información de 46.771 estudiantes (considerando la cantidad de adolescentes promedio por grupo en cada tipo de centro). Se seleccionaron como variables de interés las habilidades en lectura y matemática, así como índices de habilidades socioemocionales y características sociodemográficas para considerar diferentes tipos de variables. Las variables seleccionadas fueron las siguientes:

- habilidad en matemática (theta_MAT_300_50),
- habilidad en lectura (theta_LEN_300_50),
- índice de motivación y autorregulación (MOTAUTREGA),
- índice de regulación emocional (REGEMO),
- índice de estatus socioeconómico y cultural de la familia (ESCS_Alumno) y
- género del alumno (AlumnoGenero).

La simulación de las variables continuas se realizó por separado para cada estrato del diseño muestral utilizando los parámetros estimados con datos de Aristas Media 2022. En cada estrato se calcularon la media y la matriz de covarianza, a fin de preservar tanto los niveles promedio como la estructura de correlaciones entre variables. A partir de estos, la generación de los datos continuos se implementó mediante la función mvrnorm del paquete MASS.

Utilizando el índice ESCS_Alumno, se derivaron dos nuevas variables: ESCS_Centro, calculado como el promedio por centro, y quintiles del índice, tanto a nivel del estudiante (ESCS_Alumno_cat) como del centro (ESCS_Centro_cat).

La variable dicotómica género[3] se simuló en varias etapas, para mantener diferencias en la habilidad en lectura y matemática en coherencia con la tendencia observada en ediciones anteriores de Aristas y en otros estudios regionales[4]. En primer lugar, se definió el número de estudiantes varones y mujeres de modo de mantener una proporción cercana a 50/50 en la población total. Para luego poder evaluar diferencias de habilidad por género, se introdujo un sesgo en lectura, ordenando a la población de mayor a menor según dicha habilidad. En el 50% superior se asignó un 54% de mujeres y un 46% de varones, mientras que en el 50% inferior se aplicó la distribución complementaria, manteniendo el balance global. Este procedimiento generó una leve ventaja promedio de las mujeres en lectura. De manera análoga, se aplicó un sesgo en matemática, ordenando la población de acuerdo con esa habilidad. En el 50% superior se asignó un 55% de varones y un 45% de mujeres, y en el 50% inferior se aplicó la distribución complementaria. Esto resultó en una ligera ventaja promedio de los varones en matemática. Finalmente, se combinaron las asignaciones de género obtenidas a partir de lectura y matemática. Cuando ambas coincidían, se conservó el género asignado. En los casos en que diferían, se asignaba aleatoriamente la categoría mujer o varón con igual probabilidad. Como resultado final, la variable género mantiene aproximadamente un 50% de mujeres y un 50% de varones. Además, las medias por género reflejan pequeñas diferencias en las habilidades por área: las mujeres tienden a presentar un desempeño promedio superior en lectura, mientras que los varones muestran una ventaja en matemática.

[3] Si bien la pregunta por género realizada en Aristas considera tres opciones de respuesta (femenino, masculino y otro), la categoría otro presenta muy baja prevalencia (2,6% en Aristas Media 2022). Considerando su baja presencia, por simplicidad, para efectos de los objetivos de este estudio, la población simulada consideró únicamente el género binario.

[4] En Aristas Media 2022 los varones presentaron resultados significativamente más altos que las mujeres en matemática, mientras que las mujeres lograron resultados significativamente más altos que los varones en lectura. La brecha registrada en matemática es consistente con los resultados reportados por PISA para Uruguay y los resultados de lectura con la tendencia identificada en varios estudios regionales (INEEd, 2023a).

TABLA 1
CARACTERÍSTICAS DE LA POBLACIÓN

Variable	Grupo	n / Media (desviación estándar)	%
Estudiantes	—	46.771	—
Grupos	—	1.953	—
Centros	—	586	—
Habilidad en lectura	Total	301,63 (54,88)	—
	Mujer	302,32 (54,77)	—
	Varón	300,94 (54,99)	—
Habilidad en matemática	Total	297,20 (51,99)	—
	Mujer	296,27 (51,77)	—
	Varón	298,13 (52,20)	—
Regulación emocional	—	49,04 (8,99)	—
Motivación y autorregulación	—	49,90 (9,29)	—
Género	Mujer	23.417	50,1
	Varón	23.354	49,9
Estatus socioeconómico y cultural	Muy desfavorable	9.355	20,0
	Desfavorable	9.354	20,0
	Medio	9.354	20,0
	Favorable	9.354	20,0
	Muy favorable	9.354	20,0
Tipo de curso	Liceo privado	7.822	16,7
	Liceo público	31.032	66,3
	Escuela técnica con educación básica integrada	5.741	12,3
	Escuela técnica con formación profesional básica	2.176	4,7
Regiones	Centro	2.044	4,4
	Este	5.384	11,5
	Montevideo	16.773	35,9
	Norte	5.612	12,0
	Oeste	7.368	15,8
	Sur	9.590	20,5

Fuente: elaboración propia a partir de datos simulados.

MUESTREO Y PONDERADORES

Se implementó un diseño de muestreo que se corresponde con el proceso estándar utilizado en Aristas Media (INEEd, 2020, 2023b), agregando algunas modificaciones definidas en 2025[5]. Se trata de un diseño en dos etapas: en la primera se seleccionan centros y en la segunda grupos.

Los centros se seleccionaron estratificando por tipo de curso (liceos públicos, liceos privados, escuelas técnicas con educación básica integrada, escuelas técnicas con formación profesional básica) y regiones (Montevideo, Sur, Centro, Este, Oeste, Norte[6]). Además, para los liceos públicos se consideró la estratificación por quintil de nivel socioeconómico[7]. Esto hace un total de 48 estratos: 6 para liceos privados, 6 para escuelas técnicas con educación básica integrada, 6 para escuelas técnicas con formación profesional básica y 30 (6 x 5) para liceos públicos.

En el caso de la educación técnica, que tiene dos tipos de curso, fue dividida en su parte con educación básica integrada y su parte con formación profesional básica. Así se "crean" dos escuelas técnicas independientes, cada una con su matrícula de estudiantes (correspondiente al tipo de curso). La parte con educación básica integrada del centro se incorpora al estrato de centros con educación básica integrada y la parte con formación profesional básica al estrato de centros con formación profesional básica. Nótese que una escuela técnica puede salir dos veces en el sorteo, una vez por sus estudiantes de educación básica integrada y una vez por sus estudiantes de formación profesional básica.

El tamaño de muestra se definió según la asignación de cuotas que satisfacen los siguientes criterios: 9.000 estudiantes a nivel nacional, 1.500 por cada tipo de centro, salvo en formación profesional básica (con 1.000), y 1.500 estudiantes por cada región, excepto Centro (con 1.000)[8].

La selección de centros se realizó con probabilidad proporcional al tamaño en todas las categorías de centro. Esto implica que el ponderador de cada centro se define como el inverso de su probabilidad de selección:

$$p_{k,h} = n_h * \frac{M_{k,h}}{M_h},$$

donde n_h es la cantidad de centros a seleccionar en el estrato h, $M_{k,h}$ es la matrícula del centro k dentro del estrato y M_h es el total de estudiantes en el estrato.

[5] Las modificaciones referidas se relacionan con el mecanismo de selección de los centros y la estratificación de los tipos de curso de educación técnica.
[6] La región Sur incluye Canelones y San José; la Este, Lavalleja, Maldonado, Rocha y Treinta y Tres; la Norte, Artigas, Cerro Largo, Rivera y Tacuarembó; la Oeste, Colonia, Paysandú, Río Negro, Salto y Soriano, y la Centro, Durazno, Flores y Florida.
[7] Calculado a partir de la metodología del índice de vulnerabilidad social (IVS) promedio del centro (ANEP, 2022).
[8] Estas fueron las cuotas utilizadas en el diseño muestral de Aristas Media 2025, las cuales incluyen tanto la muestra regular de la evaluación definitiva como una cantidad adicional de casos a utilizar en estudios puntuales. En el marco de este trabajo se utilizó como referencia el número total de casos. Si bien esta cifra no refleja el tamaño de la muestra de Aristas (el cual es un poco menor), esto no representa un problema a efectos de la comparación de métodos realizada en este estudio, que evalúa a todos bajo las mismas condiciones.

Para la selección de estudiantes, se aplicó el criterio de elegir al azar dos grupos por centro (si solo había un grupo, se seleccionaba ese). La probabilidad de selección de cada estudiante dentro de su centro se describe como:

$$p_{i,k,h} = \frac{m_{k,h}}{M_{k,h}} ,$$

donde $m_{k,h}$ es la cantidad de estudiantes seleccionados en el centro k del estrato h. Esta se multiplica por la probabilidad del centro y el ponderador del estudiante i del centro k en el estrato h se calcula como:

$$w_{i,k,h} = (p_{k,h} * p_{i,k,h})^{-1} .$$

Con este diseño se simularon 1.000 extracciones de muestra. Las características de las muestras generadas se presentan en la tabla 2.

TABLA 2
DESCRIPCIÓN DE LAS MUESTRAS UTILIZADAS EN LA SIMULACIÓN

Número de iteraciones	1.000
Centros	304
Distribución de centros por categoría	Liceo privado (14,5%), liceo público (56,2%), escuela técnica con educación básica integrada (14,2%), escuela técnica con formación profesional básica (15,1%)
Distribución de centros por regiones	Centro (9,9%), Este (13,8%), Montevideo (28,9%), Norte (15,5%), Oeste (14,8%), Sur (17,1%)
Grupos (media ± sd)	1.376 ± 12
Estudiantes (media ± sd)	13.350 ± 94

Fuente: elaboración propia a partir de datos simulados.

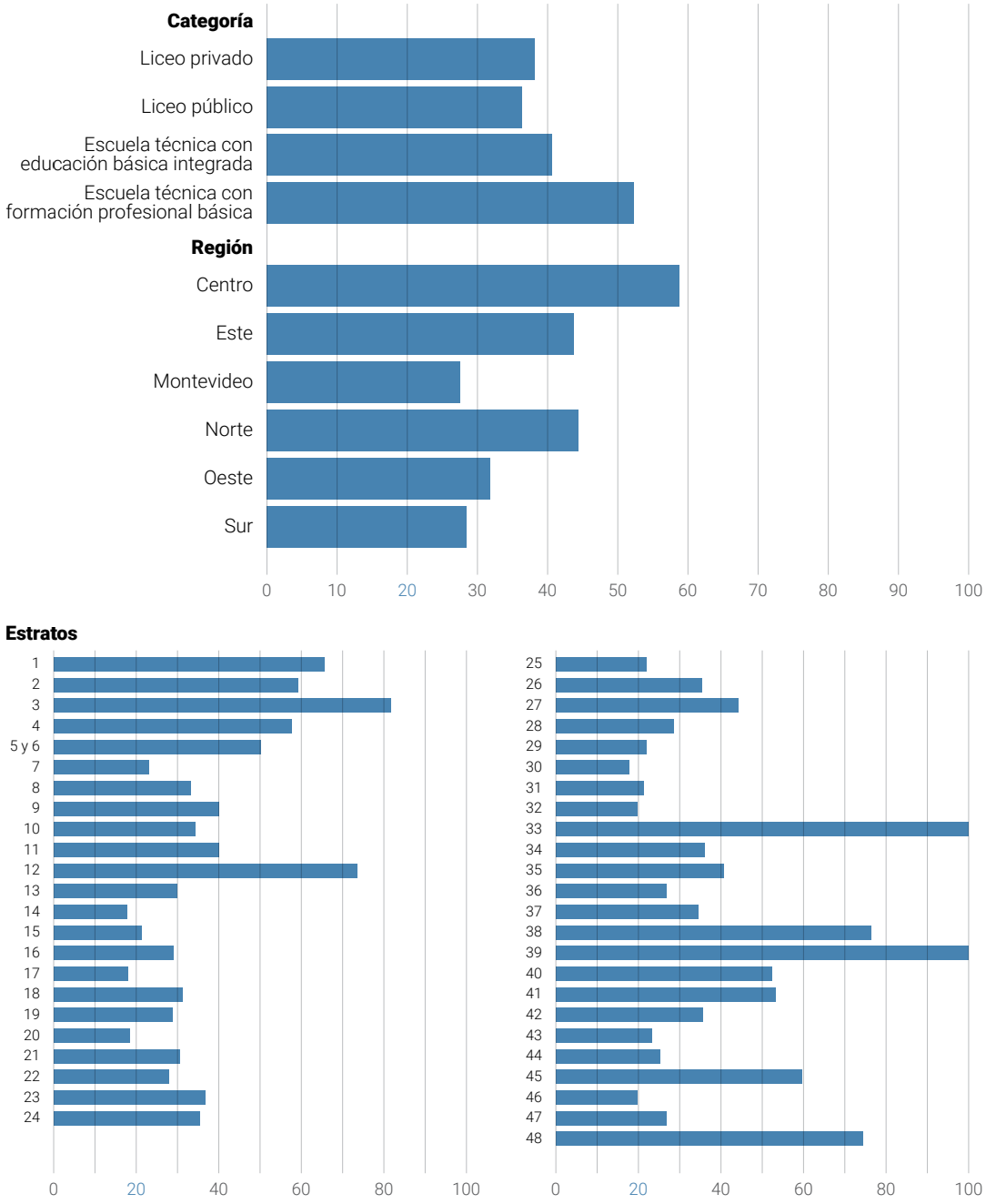
Para simplificar el análisis se asume que todos los estudiantes fueron evaluados, por lo cual no se aplicaron ajustes por no respuesta a los ponderadores. Asimismo, tampoco se realizan modificaciones de la escala de los ponderadores.

FRACCIONES DE MUESTREO

El gráfico 1 presenta en el panel superior las fracciones de muestreo promedio según tipo de centro y región, calculadas como la cantidad de estudiantes seleccionados en la muestra sobre la cantidad de estudiantes en la población, tomando como referencia una de las múltiples muestras extraídas. Se observa que la tasa promedio es superior entre los estudiantes de formación profesional básica respecto a los de educación básica integrada (en liceos públicos, privados y escuelas técnicas) y en las regiones Centro, Este y Norte del interior del país. En el panel inferior se muestran las fracciones de muestreo en cada estrato explícito del diseño muestral.

GRÁFICO 1

FRACCIÓN DE MUESTREO PROMEDIO SEGÚN CATEGORÍA Y REGIONES Y POR ESTRATO EN MUESTRA 1



Fuente: elaboración propia a partir de datos simulados.
Nota: para la estimación de varianza el estrato 5 se colapsa con el 6 debido a que contiene solo un centro educativo.

ESTIMACIÓN DE LA VARIANZA

En estudios basados en muestras complejas, como es el caso de Aristas, la estimación de la varianza de los estimadores requiere técnicas que aproximen lo que sería el cálculo directo mediante fórmulas explícitas, ya que este cálculo directo suele no ser viable.

Se exploraron distintos métodos de estimación de la varianza, que incluyen aproximaciones basadas en fórmulas y métodos de réplica. Cada método tiene propiedades y limitaciones específicas en función del diseño de muestreo y del tipo de estimador (medias, cuantiles, etc.) que se desea analizar. La comparación entre estos enfoques permite evaluar el impacto de la corrección por población finita y de la estructura del diseño en la precisión de los estimadores.

DISEÑO CON Y SIN FACTOR DE CORRECCIÓN PARA POBLACIONES FINITAS (FPC)

Särndal et al. (1992) presentan el V_0, una aproximación al cálculo de la varianza que se basa en muestras realizadas con reemplazo. El $V_{0,h}$ para un estrato h para un estimador $\hat{\theta}_h$ de promedios o totales es:

$$V_{0,h} = (n\times(n-1))^{-1} \Sigma_{S_h} (n/p_{k,h}\times\hat{\theta}_{k,h}-\hat{\theta}_h)^2 \ ,$$

donde $\hat{\theta}_{k,h}$ es la estimación del parámetro en el centro k de ese estrato y $\hat{\theta}_h$ es la estimación en el estrato h. Para cuantiles, el comando svyquantile() de R tiene un procedimiento similar[9].

En el caso de que la fracción de muestreo sea no despreciable[10], como es el caso en la mayoría de los estratos de Aristas Media, el estimador de $V(\hat{\theta}_h)$ es:

$$(1-f_h)V_{0,h} \ ,$$

donde f_h es la fracción de muestreo en el estrato y $(1-f_h)$ es el factor de corrección para poblaciones finitas (f_{pc} por sus siglas en inglés) (Valliant et al., 2018). Para el total de la muestra el estimador de $V(\hat{\theta})$ es:

$$\Sigma_{h=1}^{H} V(\hat{\theta}_h) .$$

Se probaron tres versiones del *fpc*:

1. $f_{1,h}=m_h/M_h$ donde m_h es la cantidad de estudiantes seleccionados en los centros de la muestra del estrato h($m_h=\Sigma_{Sh}\ m_{k,h}$) y M_h es la matrícula total en ese estrato ($M_h= \Sigma_k\ M_{k,h}$).

2. $f_{2,h}=n_h/N_h$ donde n_h es la cantidad de centros seleccionados de los N_h centros de ese estrato.

3. $f_{3,h}=M_{sh}/M_h$ donde M_{sh} es la matrícula total de los centros seleccionados en el estrato h y M_h es la matrícula total en ese estrato ($M_{sh}= \Sigma_{Sh}\ m_{k,h}$).

[9] Ver documentación en el paquete survey (Lumley, 2004).
[10] Valliant et al. (2018, p. 95) refiere a tasas superiores al 5%.

Se observa que $f_{3,h}$>=$f_{2,h}$, mientras que $f_{2,h}$ es mayor que $f_{1,h}$ en la mayoría de los estratos, por lo tanto, utilizar $f_{2,h}$ o $f_{3,h}$ en el cálculo de la varianza implica una corrección mayor por población finita que con $f_{1,h}$.

También se incorporó en el análisis el cálculo de la varianza "sin *fpc*", es decir, el estimador de $V(\hat{\theta})$ es en este caso $\Sigma_{h=1}^{H} V_{0,h}$ sin considerar los factores de corrección por población finita.

BRR-FAY

El método BRR fue creado en el Bureau del Censo en Estados Unidos para el caso de diseños con muchos estratos y dos elementos (*primary sampling units*, PSUs) en cada estrato (McCarthy, 1966). Originalmente se formaban *R* medias muestras, un elemento por estrato, de tal forma que fuesen balanceadas en la cantidad de veces que cada PSU termina apareciendo en las muestras. Este balance se logra mediante el uso de matrices ortogonales, conocidas como matrices de Hadamard, que indican en cada caso cuál de las PSU va a ser elegida en cada uno de los estratos. Para cada una de las *R* medias muestras, un estimador $\hat{\theta}_b$ se calcula y la $V(\hat{\theta})$ se estima con:

$$\frac{1}{R} \times \Sigma_{b=1}^{R} (\theta_b - \theta)^2.$$

En la variante de Fay (1989), en lugar de considerar una sola de las PSUs duplicando su ponderador, se utilizan ambas unidades haciendo una combinación lineal con un factor de perturbación del 50%[11]: se multiplica por 1,5 al ponderador de una de las PSUs y 0,5 al ponderador de la otra (en el caso de ternas se multiplica por 1,7071 a una PSU y 0,6464 a las otras dos).

Dado que el diseño en Aristas Media no es de dos elementos por estrato (PSUs), se ordenan los centros por matrícula dentro de cada estrato y se forman pseudoestratos colocando dos centros en cada uno. En los casos que n_h sea impar, el último pseudoestrato contiene una terna de centros. Se utiliza una matriz de Hadamard de 160 x 160, resultando en 160 réplicas para la estimación de la varianza. Las réplicas deben ser calibradas para que la suma de los ponderadores en cada estrato reproduzca la matrícula del estrato.

El método BRR-Fay tiene la característica de proveer estimadores de varianza legítimos para estadísticos lineales y no lineales, incluyendo cuantiles, los cuales no son bien estimados por otras metodologías de replicación como el JK2 (Valliant et al., 2018).

Un punto importante, y que fue clave para realizar este estudio, es que el método BRR-Fay no permite la inclusión de factores de corrección por población finita. Por esta razón, su desempeño se contrasta con otros enfoques que sí incorporan *fpc*, permitiendo evaluar el impacto de esta corrección en la estimación del error muestral.

En el Anexo se presentan los comandos en R y Stata para la implementación de estos métodos.

[11] Fay (1989) propuso un factor de perturbación del 50%. Estudios posteriores han evaluado el uso de otros factores (10%, 30%) (Judkins, 1990).

PARÁMETROS DE INTERÉS ESTIMADOS

Se aplicó el estimador Horvitz-Thompson para calcular los parámetros de interés a partir de las 1.000 muestras simuladas. Este es un método de estimación ampliamente utilizado en encuestas que provee estimadores insesgados para estadísticas lineales y no lineales, y toma en cuenta probabilidades de selección desiguales (Valliant et al., 2018).

Se estimaron indicadores reportados usualmente en los análisis con Aristas, incluyendo medias y percentiles. Las medias se calcularon tanto a nivel global como a nivel de subgrupos (por ejemplo, por género, nivel socioeconómico, región), lo que permitió además estimar diferencias de medias entre subgrupos. Los percentiles, como la mediana (50), el 25 y el 75, se estimaron utilizando distribuciones acumuladas ponderadas, también a nivel global y por subgrupos.

COMPARACIÓN DE MÉTODOS

La consistencia de cada método en el conjunto de muestras se evalúa a través de tres medidas: cobertura (Judkins, 1990), sesgo relativo y estabilidad del error de estimación (Atasever et al., 2025).

La **cobertura** se define como la proporción de veces que el intervalo de confianza al 95% calculado a partir del estimador de varianza incluye el valor verdadero del parámetro (estimado con los datos de la población).

$$Cobertura = \frac{1}{M} \sum_{m=1}^{M} I(\Theta \in IC_m^{95\%})$$

donde M = 1.000 es el número de muestras, $IC_m^{95\%}$ es el intervalo de confianza en la muestra m. Una cobertura cercana a 0,95 indica que el estimador de varianza produce intervalos de confianza insesgados.

Para calcular la cobertura en una prueba de diferencia de medias, la diferencia de medias estimada $\hat{d}_m$ en cada muestra se compara con la diferencia observada en la población d. Se calcula el sesgo en la muestra como: $\Delta_m = \hat{d}_m - d$ y se analiza la proporción de veces en que el intervalo de confianza al 95% de Δ_m incluye el valor nulo (0), que indica que la diferencia encontrada en la muestra es igual a la encontrada en la población:

$$Cobertura = \frac{1}{M} \sum_{m=1}^{M} I(0 \in IC_m^{95\%})$$

Por otra parte, el "verdadero" error de estimación se aproxima mediante la desviación estándar de las estimaciones del parámetro a lo largo de las M muestras:

$$SE_{verdadero} = sd(\hat{\theta}_1, \ldots, \hat{\theta}_M)$$

Para cada estimador de varianza, el error promedio estimado se calcula como:

$$\hat{SE} = \frac{1}{M} \sum_{m=1}^{M} \hat{SE}_m$$

donde $\hat{SE} = \sqrt{\hat{var}(\hat{\theta}_m)}$ es el error estándar estimado en la muestra m.

El **sesgo relativo** se define como la razón entre el error promedio estimado y el verdadero error de estimación:

$$Sesgo\ relativo = \frac{\hat{SE}}{SE_{verdadero}}$$

Un sesgo relativo de 1 indica que el estimador aproxima correctamente el error muestral. En contraste, valores mayores a 1 indican sobreestimación y menores a 1 subestimación.

La **estabilidad** mide la variabilidad de los errores estándar estimados por un método, en relación con el verdadero error de estimación:

$$Estabilidad = \frac{sd(\hat{SE}_1, \ldots, \hat{SE}_M)}{SE_{verdadero}}$$

Este indicador se corresponde con el coeficiente de variación de la estimación del error. Valores bajos (cercanos a cero) de esta medida indican que el estimador de varianza es más consistente, indicando mayor estabilidad del método de estimación de la varianza.

RESULTADOS

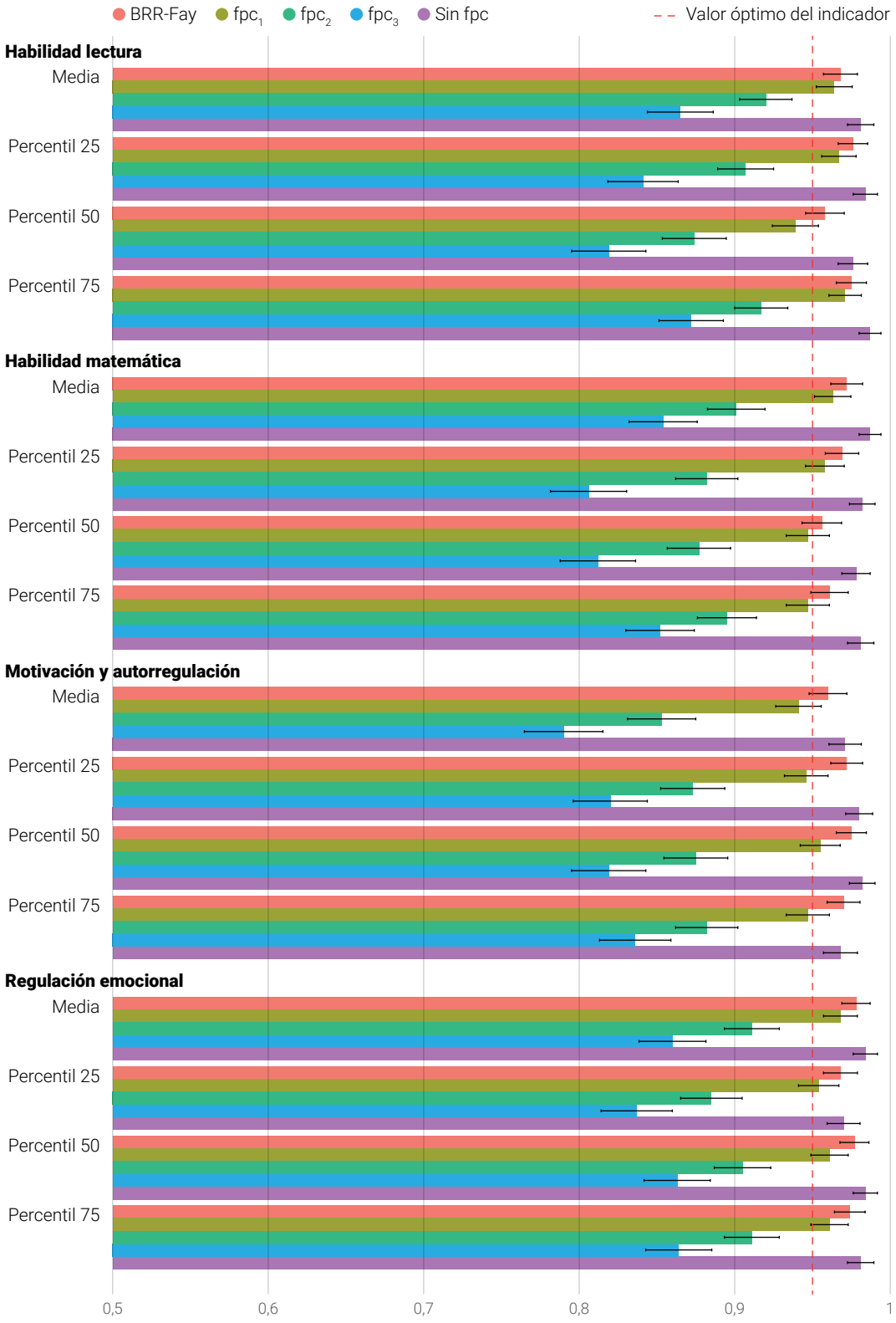
MEDIAS Y PERCENTILES EN LA MUESTRA TOTAL

En el gráfico 2 se muestra la cobertura de los intervalos de confianza para la media y los cuantiles de las variables de interés según los distintos métodos de estimación de la varianza (diseño con y sin factor de corrección para poblaciones finitas —fpc— y BRR-Fay). Se observa que con el método BRR-Fay la precisión mejora respecto al diseño sin corrección por *fpc*. Sin embargo, en general es menos preciso que el diseño con corrección por fpc_1, calculado como la proporción de estudiantes seleccionados en la muestra. Las variantes fpc_2 y fpc_3, basadas en la proporción de escuelas y la proporción de estudiantes en los centros seleccionados, que consideran una corrección por fracciones mayores a las de la muestra efectiva, presentan un peor desempeño que se aleja del nivel de cobertura esperado (95%) y, lo que es más grave, subestimando el error estándar.

El gráfico 3 muestra la distribución de los indicadores de cobertura, sesgo relativo del error estándar y estabilidad del error, considerando todas las variables y medidas evaluadas (16 pruebas en total). Respecto a la cobertura, como se mencionó anteriormente, el diseño con corrección fpc_1 presenta mayor precisión, con valores generalmente cercanos a 0,95, seguido por BRR-Fay, que tiende a sobreestimar ligeramente la cobertura. Un patrón similar se observa al analizar el sesgo relativo: fpc_1 se acerca al valor óptimo, mientras que BRR-Fay lo sobreestima levemente. En cuanto a la estabilidad del error, BRR-Fay muestra porcentajes mayores, indicando una menor estabilidad en comparación con el resto de los métodos.

GRÁFICO 2

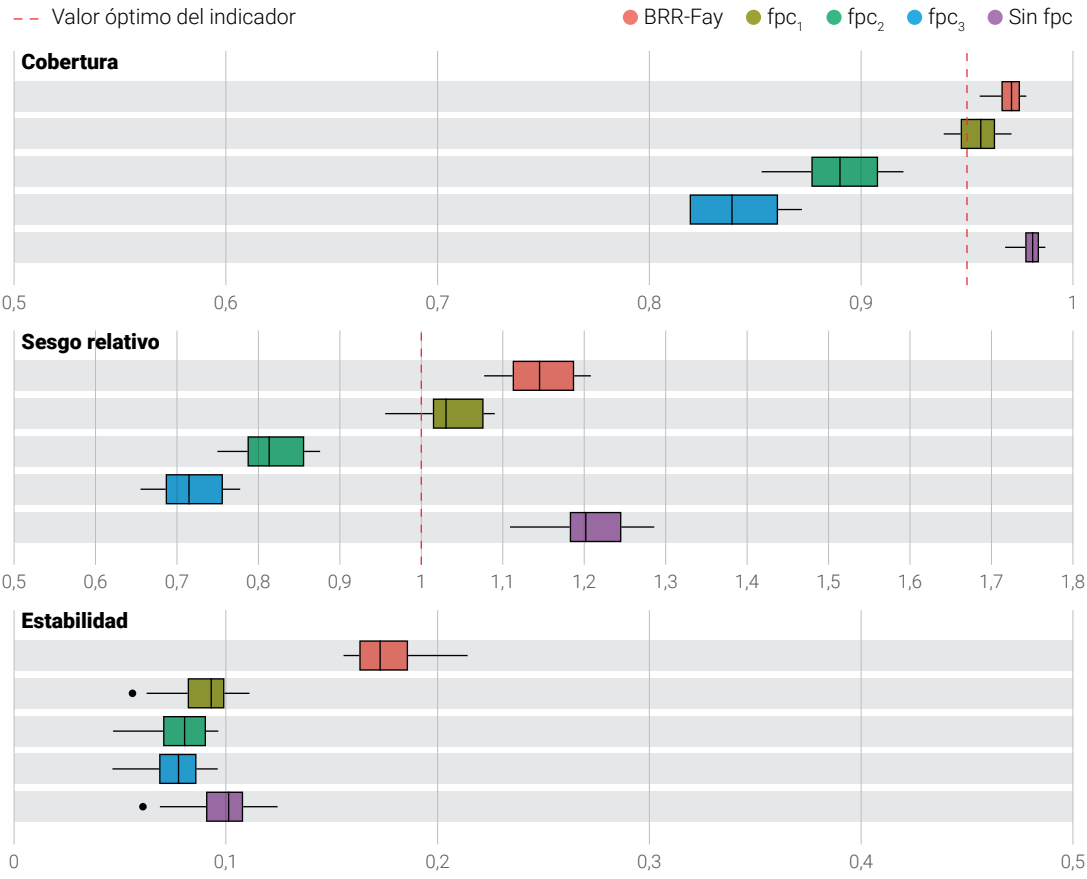
COBERTURA DEL VALOR VERDADERO (IC 95%) POR VARIABLE Y PARÁMETRO ESTIMADO SEGÚN MÉTODO DE ESTIMACIÓN DE LA VARIANZA



Fuente: elaboración propia a partir de datos simulados.
Nota: BRR-Fay, réplicas repetidas balanceadas con corrección de Fay; *fpc_1*, *fpc* igual a la proporción de estudiantes incluidos en la muestra; *fpc_2*, *fpc* igual a la proporción de centros incluidos en la muestra; *fpc_3*, *fpc* igual a la proporción del total de estudiantes pertenecientes a los centros incluidos en la muestra; *sin fpc*, sin corrección por población finita.

GRÁFICO 3

DISTRIBUCIÓN DE COBERTURA DEL IC, SESGO RELATIVO Y ESTABILIDAD DEL ERROR PARA LAS ESTIMACIONES EN EL TOTAL DE LA MUESTRA SEGÚN MÉTODO DE ESTIMACIÓN DE LA VARIANZA

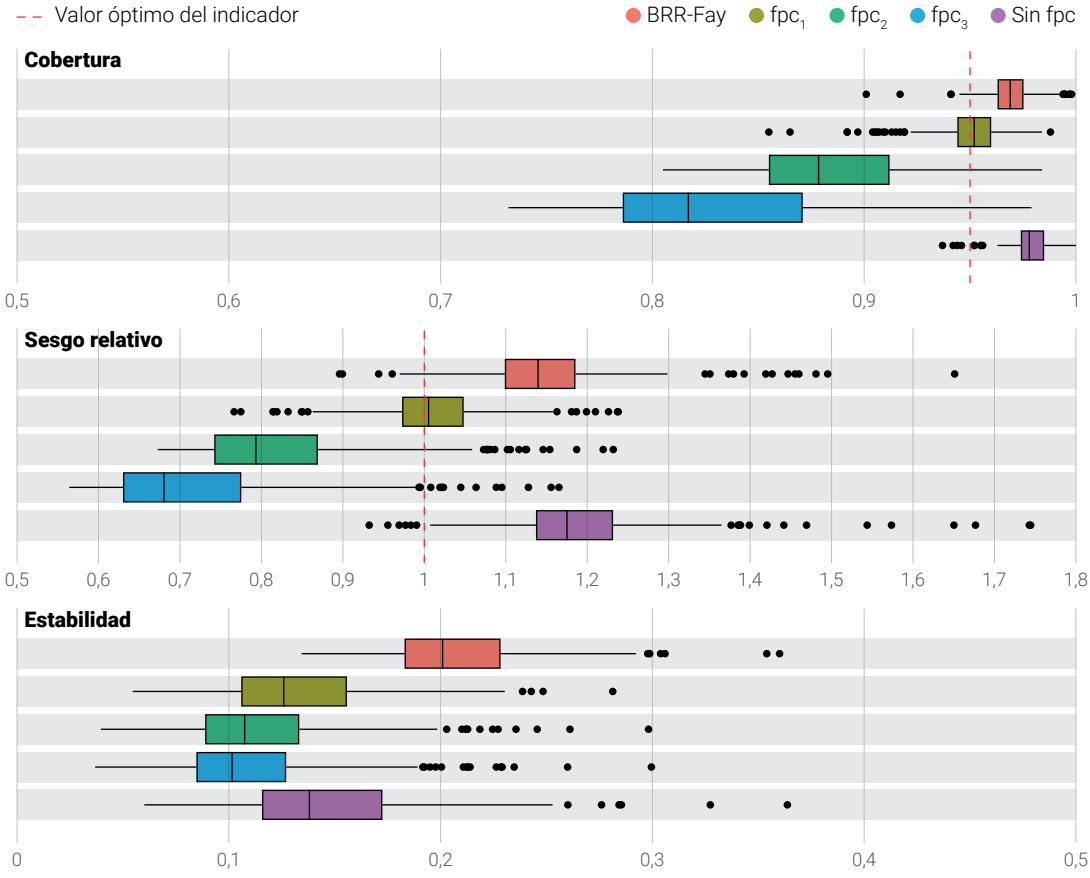


Fuente: elaboración propia a partir de datos simulados.
Nota: BRR-Fay, réplicas repetidas balanceadas con corrección de Fay; *fp,: fpc* igual a la proporción de estudiantes incluidos en la muestra; *fpc*$_2$, *fpc* igual a la proporción de centros incluidos en la muestra; *fpc*$_3$, *fpc* igual a la proporción del total de estudiantes pertenecientes a los centros incluidos en la muestra; *sin fpc*, sin corrección por población finita.

MEDIAS Y PERCENTILES EN SUBGRUPOS

Al analizar las medias y percentiles de las variables estimadas (habilidad en matemática, lectura, motivación y autorregulación, y regulación emocional) en subgrupos de la población (según las categorías de género, región, tipo de curso y quintil de contexto socioeconómico y cultural del estudiante y del centro), se contabilizaron un total de 288 pruebas. Se observa que los métodos tienen un comportamiento similar al analizado en las estimaciones en el total de la muestra: el diseño con fpc_1 presenta una cobertura y un sesgo relativo más adecuados, seguido por el método BRR-Fay. Respecto a la estabilidad, BRR-Fay tiene un desempeño inferior al de los métodos sin réplicas (gráfico 4), aunque los valores absolutos (inferiores a 0,2) se mantienen dentro de niveles aceptables.

GRÁFICO 4
DISTRIBUCIÓN DE COBERTURA DEL IC, SESGO RELATIVO Y ESTABILIDAD DEL ERROR PARA LAS ESTIMACIONES POR SUBGRUPOS SEGÚN MÉTODO DE ESTIMACIÓN DE LA VARIANZA

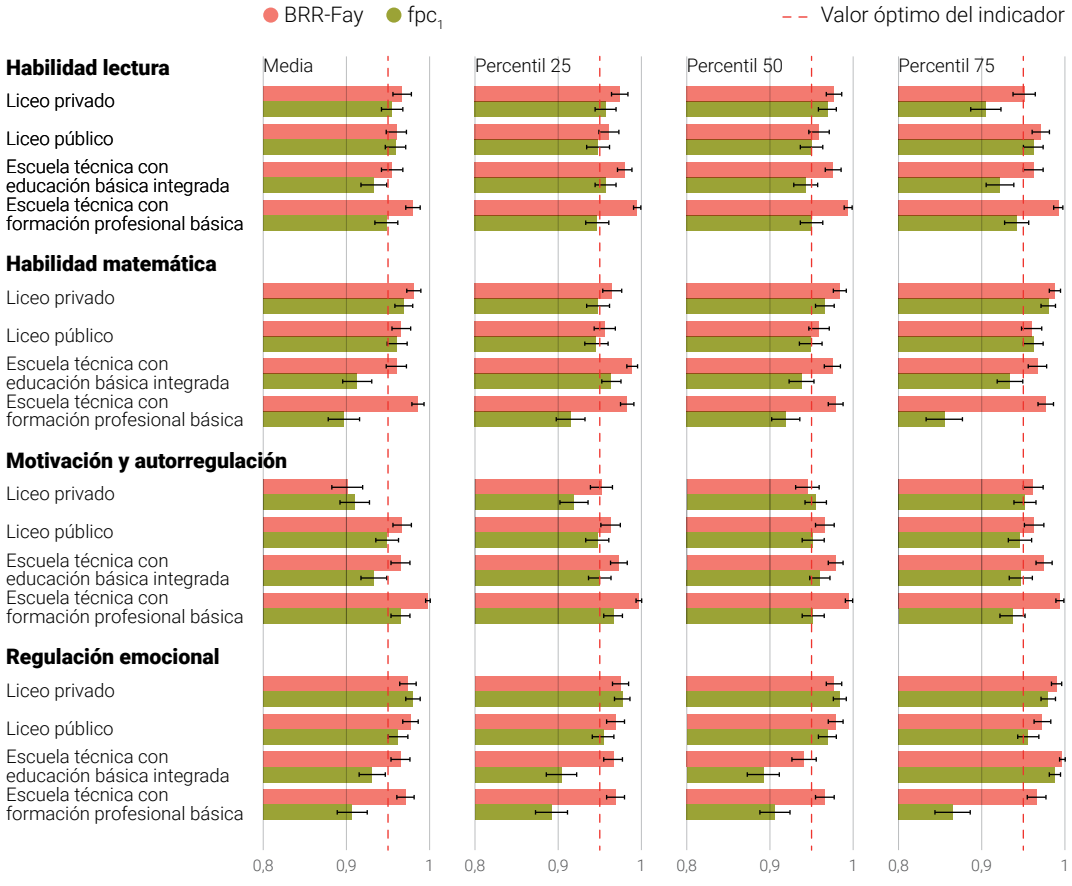


Fuente: elaboración propia a partir de datos simulados.
Nota: BRR-Fay, réplicas repetidas balanceadas con corrección de Fay; *fpc_1*, *fpc* igual a la proporción de estudiantes incluidos en la muestra; *fp_c*: *fpc* igual a la proporción de centros incluidos en la muestra; *fpc_3*, *fpc* igual a la proporción del total de estudiantes pertenecientes a los centros incluidos en la muestra; *sin fpc*, sin corrección por población finita.

COBERTURA DE IC EN CATEGORÍAS CON MAYOR FRACCIÓN DE MUESTREO

En el análisis de estimaciones en subgrupo, interesa observar las aperturas donde la fracción de muestreo es relativamente mayor, con el fin de evaluar el comportamiento de los métodos en los casos en que la corrección por *fpc* es más relevante. El gráfico 5 muestra la proporción de cobertura de las medidas estimadas según tipo de curso. Se observa que, efectivamente, en formación profesional básica, donde la tasa de muestreo es relativamente mayor que en el resto de las categorías, las diferencias en la cobertura entre los métodos son más pronunciadas. Aunque BRR-Fay tiende a sobreestimar la cobertura en formación profesional básica en todas las variables, el método de diseño con fpc_1 en algunas situaciones la subestima, especialmente en el puntaje de habilidad matemática y en el de regulación emocional.

GRÁFICO 5
COBERTURA DEL IC PARA LAS ESTIMACIONES DE MEDIAS Y PERCENTILES POR TIPO DE CURSO SEGÚN MÉTODO DE ESTIMACIÓN BRR-FAY Y DISEÑO CON fpc_1



Fuente: elaboración propia a partir de datos simulados.
Nota: BRR-Fay, réplicas repetidas balanceadas con corrección de Fay; fpc_1, *fpc* igual a la proporción de estudiantes incluidos en la muestra.

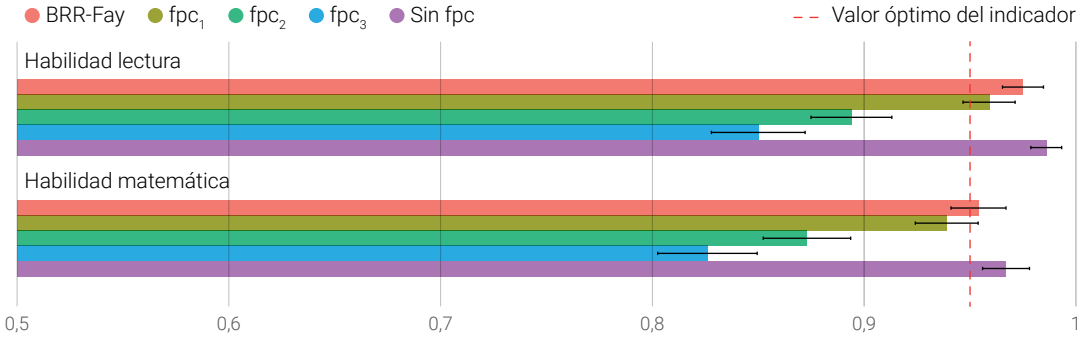
DIFERENCIA DE MEDIAS

Para evaluar el desempeño de los métodos en una prueba de diferencia de medias, se utiliza la prueba *t* para comparar los puntajes de lectura y matemática entre mujeres y varones. Tal como se mencionó en la sección de simulación, los datos poblacionales muestran una diferencia a favor de las mujeres en lectura y a favor de los varones en matemática.

El gráfico 6 presenta la proporción de cobertura de los métodos BRR-Fay y del diseño con corrección por población finita (fpc_1). Se observa que este último obtiene un desempeño adecuado en ambas variables. Sin embargo, en las estimaciones de habilidad en matemática, el método BRR-Fay muestra una cobertura más cercana al valor óptimo correspondiente con el 95%.

GRÁFICO 6

COBERTURA DEL IC PARA LAS PRUEBAS DE DIFERENCIA DE MEDIAS POR GÉNERO SEGÚN MÉTODO DE ESTIMACIÓN DE LA VARIANZA



Fuente: elaboración propia a partir de datos simulados.
Nota: BRR-Fay, réplicas repetidas balanceadas con corrección de Fay; fpc_1, *fpc* igual a la proporción de estudiantes incluidos en la muestra.

SÍNTESIS Y DISCUSIÓN

El objetivo de este estudio fue evaluar el desempeño de distintos métodos de estimación del error muestral en el marco de la evaluación Aristas. Se procura garantizar que las pruebas de significancia y los intervalos de confianza se construyan de manera adecuada, considerando el diseño muestral complejo utilizado para seleccionar a los estudiantes participantes.

La comparación se centró en el método de replicación BRR-Fay —práctica habitual en evaluaciones educativas internacionales y, en particular, en Aristas— y alternativas basadas en el diseño muestral con y sin corrección por población finita. Esta comparación cobra particular relevancia en el contexto uruguayo en Aristas Media, dado que la población no es muy grande y las fracciones de muestreo son relativamente altas, lo que hace pertinente analizar explícitamente el efecto de la corrección por población finita en la estimación de la varianza.

El análisis contempló el desempeño de cada método en la estimación de medias y cuantiles de las habilidades de lectura y matemática, así como en índices de habilidades socioemocionales. Se consideraron estimaciones tanto para el total de la muestra como para distintos subgrupos definidos por variables de contexto, área geográfica y género. Además, se incluyeron comparaciones basadas en tests para diferencias de medias.

Para evaluar el desempeño de cada método se utilizaron como indicadores principales la cobertura de los intervalos de confianza y el sesgo relativo de la varianza estimada, contrastando las estimaciones obtenidas con los valores poblacionales conocidos en la simulación. La cobertura se definió como la proporción de intervalos que incluían el valor verdadero del parámetro, mientras que el sesgo relativo reflejó la desviación del promedio de las varianzas estimadas respecto de la varianza "verdadera". Adicionalmente, se examinó la estabilidad de los resultados a través de la variabilidad de las estimaciones de las varianzas en las diferentes réplicas de la simulación.

Resultados generales

En términos generales, los resultados de la simulación muestran que el método BRR-Fay presenta un desempeño adecuado, mejorando la aproximación para muestras con reemplazo, con indicadores cercanos al valor óptimo en términos de cobertura y sesgo relativo, lo que respalda su uso continuado en la práctica. Sin embargo, el diseño con corrección por población finita con el factor fpc_1 —definido como la proporción de estudiantes seleccionados en la muestra en cada estrato— obtiene resultados más precisos que el BRR-Fay en cobertura y sesgo relativo, situando a esta metodología como una opción a considerar. Otras variantes

del factor de corrección por población finita (fpc_2 y fpc_3) que incorporan una corrección por fracción de muestreo mayor arrojaron estimaciones sesgadas, subestimando el error en todos los casos analizados, lo que descarta su pertinencia práctica.

En cuanto a la estabilidad, el diseño con fpc_1 muestra cifras más estables en las estimaciones del error para el total de casos y por subgrupos en todas las medidas y variables de interés, mientras que BRR-Fay produce estimaciones con menor estabilidad. Ahora bien, es importante señalar que los resultados de estabilidad del BRR-Fay (entre 0,15 y 0,20) son similares a los obtenidos por esta técnica en el estudio de IEA (Atasever et al., 2025) y bastante más estables que los obtenidos en el estudio de Judkins (1990) para esta misma técnica en otros estimadores. Es decir, si bien BRR-Fay presenta estimaciones menos estables que los otros dos métodos evaluados, en términos absolutos las cifras de estabilidad se encuentran dentro de rangos aceptables.

Resultados en casos particulares

Ambos métodos, BRR-Fay y el diseño con fpc_1, fueron comparados en situaciones particulares de interés. Primero, se consideraron subgrupos con fracciones de muestreo relativamente altas. En estos casos, la brecha entre la cobertura de los métodos se amplió, comparando con otros subgrupos, y se observa que BRR-Fay tiende a sobreestimar ligeramente la varianza, mientras que el diseño sin réplicas con fpc_1 mostró, en algunas variables, una subestimación mayor, es decir, un sesgo más pronunciado que el BRR-Fay.

En segundo lugar, se analizó la cobertura de la prueba *t* para la diferencia de medias por género en habilidades de lectura y matemática. Allí se observó un patrón similar: BRR-Fay presentó una ligera sobreestimación para detectar diferencias en ambas habilidades, mientras que la corrección por fpc_1 fue más precisa en la prueba de lectura, pero en matemática subestimó. En este último caso, la subestimación del diseño con fpc_1 es mayor que la sobreestimación que se obtuvo con el BRR-Fay.

Estos resultados indican que, si bien mediante la aplicación de fpc_1 se obtuvieron estimaciones más precisas y estables del error en términos generales, hay indicios de que, en situaciones con alta fracción de muestreo o en ciertas variables específicas, podría mostrar un peor desempeño que el BRR-Fay, subestimando la varianza. Por lo anterior, el fpc_1 podría considerarse un método menos conservador.

Discusión

Los resultados obtenidos deben interpretarse considerando el propósito inferencial del estudio. Como se señala en OCDE (2002), incluso cuando se dispone de información de toda la población, la estimación de la varianza sigue siendo necesaria. PISA no aplica correcciones por población finita, ya que considera infinita la población de escuelas y de estudiantes (OCDE, 2024). En este sentido, en el caso uruguayo, resulta relevante distinguir si el objetivo es describir únicamente a los 46.771 individuos específicos incluidos en el análisis o si se quiere realizar inferencia respecto de la superpoblación teórica que generó dicha población. En este último caso, la aplicación de una corrección por población finita podría

no corresponder, dado que el interés inferencial se orienta a una población teórica más amplia. En este caso, su aplicación podría reducir artificialmente la varianza, subestimando la incertidumbre de las estimaciones.

El uso de BRR-Fay como método de estimación de varianza está ampliamente documentado en evaluaciones educativas internacionales de gran escala (PISA, TALIS, TIMSS, PIRLS, entre otras). En el presente estudio, los resultados obtenidos con este método fueron consistentes y satisfactorios, confirmando su idoneidad para la estimación de varianza en Aristas. Además, su aplicación asegura continuidad con ediciones anteriores, facilitando la comparabilidad de resultados, lo que constituye un aspecto clave. Asimismo, la elección de un método de replicación como BRR-Fay se asocia también a consideraciones de confidencialidad y resguardo de bases de microdatos, aspecto relevante para el contexto de Aristas.

Por otra parte, el método basado en el diseño con corrección por población finita (fpc_1) también mostró un desempeño adecuado, por lo que puede considerarse una alternativa válida. Esta metodología presenta ventajas en términos de simplicidad, precisión y eficiencia computacional, por lo que podría considerarse en análisis secundarios o modelos estadísticos específicos. No obstante, su carácter menos conservador en algunos escenarios requiere una evaluación cuidadosa.

Aportes y limitaciones

El presente reporte aporta evidencia sobre el desempeño de distintas metodologías de estimación de varianza en el marco de Aristas, contribuyendo al fortalecimiento de las evaluaciones educativas en Uruguay. Al mismo tiempo, los hallazgos pueden resultar de interés para otros países de la región que enfrentan desafíos similares en el uso de diseños muestrales complejos y en la pertinencia de aplicar correcciones por población finita en contextos con poblaciones pequeñas y fracciones de muestreo no despreciables.

Como limitación, se debe señalar que las pruebas realizadas se restringieron a un conjunto de estimadores y variables específicas. Asimismo, se trabajó con un tamaño fijo de centros en las muestras y bajo el supuesto de inexistencia de no respuesta. Futuras exploraciones podrían contemplar otras medidas, como correlaciones o coeficientes de regresión, otras variables y el levantamiento de supuestos, así como, eventualmente, la comparación con otros métodos de estimación de varianza.

Conclusión

En síntesis, el método BRR-Fay muestra un desempeño adecuado y puede ser preferible en Aristas por motivos de comparabilidad, confidencialidad y por ser un enfoque más conservador. No obstante, la corrección por población finita mediante fpc_1 emerge como alternativa válida que generalmente mejora la precisión de las estimaciones.

Por lo tanto, ambos métodos pueden considerarse recomendables. En caso de que no existan restricciones de identificación u otros inconvenientes, la información sobre el diseño podrá

ser puesta a disposición de terceros junto con los datos publicados para su uso en el análisis de Aristas.

ANEXO

A continuación, se detallan los comandos de R y Stata para el análisis de los datos de Aristas Media aplicando los métodos de estimación de varianza BRR-Fay (0,5) y el diseño sin réplicas con corrección para poblaciones finitas.

1) BRR-Fay (0,5)
R (srvyr):

```
design_brr <- dat_dis %>%
    as_survey_rep(
        weights = peso_MEst,
        repweights = starts_with("W_REP_"),
        type = "Fay",
        combined_weights = TRUE,
        rho = 0.5,
        mse = FALSE)
```

Stata:

```
svyset [pweight = peso_MEst], brrweight(W_REP_*) fay(0.5)
```

2) Diseño sin réplicas con factor de corrección fpc1.
R (srvyr):

```
design_fpc1 <- dat_dis %>%
    as_survey(ids = CentroCodigo,
    strata = estrato2,
    nest = TRUE,
    fpc = fpc1,
    weights = peso_MEst)
```

Stata:

```
svyset CentroCodigo [pweight = peso_MEst], strata(estrato2) fpc(fpc1)
```

REFERENCIAS BIBLIOGRÁFICAS

ANEP. (2022). *Índice de Vulnerabilidad Socioeconómica en Enseñanza Media. Diciembre 2022*. https://observatorio.anep.edu.uy/sites/default/files/arch/IVSEducMediaANEP_InformeMetodologico202303.pdf

Atasever, U., Meinck, S. y Cortes, D. (2025). *An Examination of the Performance of Variance Estimators in International Large-Scale Assessments*. https://www.iea.nl/sites/default/files/2025-07/RD-Call-Three-Variance-Estimators.pdf

Fay, R. E. (1989). *Theory and Application of Replicate Weighting for Variance Calculations*. 212–217. http://www.asasrms.org/Proceedings/papers/1989_033.pdf

INEEd. (2020). *Aristas 2018. Informe de resultados de tercero de educación media*. INEEd. https://www.ineed.edu.uy/images/Aristas/Publicaciones/Aristas2018/Aristas-2018-Informe-de-resultados.pdf

INEEd. (2023a). *Aristas 2022. Informe de resultados de tercero de educación media*. https://www.ineed.edu.uy/images/Aristas/Publicaciones/Aristas2022/Aristas-2022-Informe-resultados-tercero-educacion-media.pdf

INEEd. (2023b). *Informe técnico de Aristas Media 2022*. https://www.ineed.edu.uy/images/Aristas/Publicaciones/Aristas2022/Informe-tecnico-Aristas-Media-2022.pdf

Judkins, D. R. (1990). Fay's Method for Variance Estimation. *Journal of Official Statistics*, *6*(3), 223–239.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, *9*(8), 1–19. https://doi.org/10.18637/jss.v009.i08

McCarthy, P. J. (1966). *Replication. An Approach to the Analysis of Data From Complex Surveys* (14; Vital and Health Statistics Series).

OCDE. (2002). *PISA 2000 Technical Report*. https://www.oecd.org/content/dam/oecd/en/publications/reports/2002/12/programme-for-international-student-assessment-pisa_g1gh2e07/9789264199521-en.pdf

OCDE. (2024). *PISA 2022 Technical Report*. https://doi.org/10.1787/01820d6d-en

Särndal, C.-E., Swensson, B. y Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.

UNESCO. (2015). *Estudio Regional Comparativo y Explicativo (ERCE 2019): reporte técnico*. https://unesdoc.unesco.org/ark:/48223/pf0000394282

Valliant, R. y Dever, J. A. (2018). *Survey Weights: A Step-by-Step Guide to Calculation* (1.ª ed.). Stata Press.

Valliant, R., Dever, J. A. y Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples* (2.ª ed.). Springer.