

Evaluación de Impacto del
Programa Todos a Aprender (PTA)
En el marco del Contrato de
Préstamo BIRF 9360-CO

Contrato interadministrativo
COI.PCCNTR.9012747 de 2026, entre
el **Ministerio de Educación Nacional**
y la **Universidad de Nariño**



**Evaluación de Impacto del Programa Todos a Aprender (PTA)
En el marco del Contrato de Préstamo BIRF 9360-CO**

Producto:

Informe de Evaluación de Impacto

**UNIVERSIDAD DE NARIÑO
Equipo Evaluador
San Juan de Pasto, mayo 13 de 2026**

Ministerio de Educación Nacional

José Daniel Rojas Medellín	Ministro
Lucy Maritza Molina Acosta	Viceministra
Solman Yamile Díaz Ossa	Directora de Calidad
Marcela Cascavita	Asesora del despacho, encargada del Programa de Tutorías para el Aprendizaje y la Formación Integral PTAFI 3.0
Mónica Figueroa Dorado	Equipo Pedagógico Nacional PTAFI 3.0

Universidad de Nariño

Martha Sofía González Insuasti	Rectora
Hernán Modesto Rivas Escobar	Decano de la Facultad de Educación
Jesús Orlando Gracia Fajardo	Director
Jairo de los Ríos	Coordinador Operativo
Luis Alberto Montenegro	Experto Cualitativo
Mónica Esmeralda Vallejo	Experta Cualitativa
Alba Patricia Palacios Valencia	Experta Temática
Carlos Alberto Meriño	Experto Temático
John Jairo Narváez Galviz.	Experto Cuantitativo
John Barco Jiménez	Experto Cuantitativo
Aura María Meneses Zamora	Equipo de Sistematización
Christian David Chamorro	Equipo de Sistematización

Contenido

Resumen ejecutivo	6
Listado de siglas y acrónimos.....	7
Índice de tablas.....	11
Índice de figuras	13
1. Introducción	17
1.1. Propósito del informe	17
1.2. Alcance de la evaluación de impacto	18
1.3. Resultados causales priorizados.....	19
1.4. Relación con la teoría de cambio del programa.....	20
2. Fundamento conceptual y causal de la evaluación.....	22
2.1. Teoría de cambio del PTAFI 3.0	22
2.2. Hipótesis causales	24
2.3. Razonamiento contrafactual.....	26
2.4. Complementariedad entre impacto cuantitativo e impacto cualitativo basado en teoría	27
3. Preguntas de evaluación de impacto.....	29
3.1. Efecto del programa sobre la fluidez lectora - EGRA (Fase II)	30
3.2. Efecto del programa sobre las tasas de eficiencia interna.....	32
3.3. Efecto del programa en dimensiones de actores educativos.....	33
3.4. Rol mediador de prácticas docentes y liderazgo directivo	35
4. Diseño metodológico cuantitativo de impacto	37
4.1. Diseño cuasiexperimental general	37
4.2. Construcción de grupos de tratamiento y control.....	38
4.3. Matching por score de propensión y soporte común	39
4.4. Estimador de doble robustez AIPW/DR	40
4.5. Estructura multinivel jerárquica (HLM) e inferencia cluster-robusta (WCB).....	47
4.6. Supuestos de identificación causal y pruebas de validez	49
4.7. Análisis de heterogeneidad - Efectos condicionales (CATE).....	51
4.8. Diagnósticos de robustez y especificaciones alternativas	53

5. Fuentes de información y bases analíticas.....	55
6. Caracterización de la muestra de impacto	60
7. Impacto en aprendizajes (EGRA Fase I y Fase II)	64
8. Impacto en eficiencia interna y trayectorias escolares	78
8.1.5. Contraste metodológico UDENAR vs. PROMISE	86
9. Impacto en actores educativos y dimensiones intermedias	95
9.1. Directivos docentes	96
9.4. Estudiantes.....	102
10. Evaluación cualitativa de impacto basada en teoría.....	110
11. Integración de hallazgos de impacto	124
12. Alcances y limitaciones del análisis causal	132
13. Conclusiones preliminares de la evaluación de impacto	138
14. Recomendaciones de política y escalamiento.....	142
15. Anexos.....	145
Referencias.....	147

Índice de tablas

Nombre de la tabla	Página
Tabla 1-1. Marco contractual del PTAFI 3.0	17
Tabla 3-1. Correspondencia entre preguntas causales, preguntas orientadoras de la Matriz de Consistencia e indicadores cuantitativos	36
Tabla 4-1. Especificación del estimador AIPW/DR	38
Tabla 4-2. Pruebas de validez del supuesto de identificación causal	41
Tabla 5-1. Bases analíticas integradas en la evaluación	55
Tabla 5-2. Variables construidas a partir de SIMAT y EFICIENCIA_EE	56
Tabla 5-3. Estructura general de los cinco instrumentos del operativo Fase II	57
Tabla 5-4. Coeficientes Alfa de Cronbach por dimensión e instrumento	58
Tabla 5-5. Tabla de validación cruzada por dimensión	59
Tabla 5-6. Distribución de la muestra por condición y zona	60
Tabla 6-1. Caracterización de la muestra EGRA Fase II	61
Tabla 6-2. Caracterización de la muestra de establecimientos educativos	62
Tabla 6-3. Caracterización de la muestra de actores educativos	63
Tabla 7-1. Estimación del ATE EGRA Fase II (AIPW/DR + WCB B=1.000)	64
Tabla 7-2. Análisis de sensibilidad: dos especificaciones del modelo Fase II	65
Tabla 7-3. Diferencias brutas preajuste EGRA y EGMA Fase I	69
Tabla 7-4. Resultado principal - ATE AIPW/DR Fase I	70
Tabla 7-5. Comparación: diferencia simple vs. ATE ajustado Fase I	71

Tabla 7-6. Coeficiente de correlación intraclase Fase I	72
Tabla 7-7. ATE EGRA por subprueba Fase I	73
Tabla 7-8. ATE EGMA por subprueba Fase I	74
Tabla 7-9. Contraste de los hallazgos AIPW/DR sobre EGRA entre Fase I y Fase II	76
Tabla 8-1. Brecha estructural y focalización rural/PDET del PTAFI 3.0	79
Tabla 8-2. Diagnóstico AIPW/DR sobre TEI	83
Tabla 8-3. Contraste metodológico UDENAR vs. PROMISE	85
Tabla 8-4. Brechas T-C 2017-2025 bajo tres especificaciones metodológicas	88
Tabla 8-5. Serie 2017-2025 de los indicadores PROMISE PDO1, IRI5, IRI17	90
Tabla 9-1. Descriptores muestrales del operativo Fase II y diagnósticos AIPW/DR	95
Tabla 9-2. ATE por dimensión en Directivos Docentes	97
Tabla 9-3. ATE por dimensión en Docentes	99
Tabla 9-4. Medias y dispersión por dimensión en Tutores	101
Tabla 9-5. ATE por dimensión en Estudiantes	103
Tabla 9-6. ATE por dimensión en Acudientes	105
Tabla 9-7. Síntesis cuantitativa de efectos AIPW/DR por actor	107
Tabla 10-1. Matriz CMO (Contexto-Mecanismo-Resultado) del PTAFI 3.0	110
Tabla 10-2. Niveles de intensidad de la evaluación cualitativa	120
Tabla 11-1. Convergencia entre inferencia causal cuantitativa y explicación cualitativa	125
Tabla 11-2. Cinco mediadores de la cadena causal del PTAFI 3.0	127
Tabla 11-3. Contexto-Mecanismo-Resultado	129

Índice de figuras

Nombre de la figura	Página
Figura 1-1. Articulación PTAFI 3.0 dentro del Programa PROMISE	18
Figura 4-1. Diagrama del flujo metodológico AIPW/DR	39
Figura 6-1. Distribución geográfica de la muestra Fase II	61
Figura 7-1. Diferencias estandarizadas de medias en covariables de balance Fase I	63
Figura 7-2. SMD entre los grupos tratado y control	63
Figura 7-3. Diagnósticos del modelo de propensión Fase I	64
Figura 7-4. Forest plot del ATE EGRA Fase II	65
Figura 7-5. Heterogeneidad del efecto AIPW/DR Fase II (CATE por subgrupos)	66
Figura 7-6. Curva ROC del modelo de propensión Fase II	67
Figura 7-7. Love plot del balance pre vs post IPW Fase II	68
Figura 7-8. Overlap del propensity score Fase II	69
Figura 7-9. Distribución de pesos AIPW Fase II	70
Figura 7-10. Wild Cluster Bootstrap del ATE EGRA Fase II	71
Figura 7-11. Validación cruzada 50x del modelo AIPW/DR Fase II	72
Figura 7-12. Distribución de puntajes EGRA y EGMA Fase I	73
Figura 7-13. Forest plot de los ATE EGRA y EGMA Fase I	74
Figura 7-14. Intervalos de confianza al 95% por subprueba EGRA Fase I	75
Figura 7-15. Intervalos de confianza al 95% por subprueba EGMA Fase I	76
Figura 7-16. Diferencias de línea base EGMA Fase I	76
Figura 7-17. Comparación visual de los tres estimadores Fase II	77

Figura 8-1. Evolución de la focalización PTAFI hacia rurales y PDET	79
Figura 8-2. Convergencia de la brecha Tratados-Controles en zonas vulnerables	80
Figura 8-3. Tasa de aprobación por género 2022-2024	81
Figura 8-4. Tasa de aprobación nacional 2017-2025	82
Figura 8-5. Tasa de aprobación por zona 2017-2025	83
Figura 8-6. Tasa de reprobación por género 2017-2024	84
Figura 8-7. Tasa de reprobación nacional 2017-2025	85
Figura 8-8. Tasa de reprobación por zona 2017-2025	86
Figura 8-9. Tasa de deserción por género 2022-2024	87
Figura 8-10. Trayectoria de la tasa de deserción 2017-2025	88
Figura 8-11. Tasa de deserción por zona 2017-2025	89
Figura 8-12. Comparación de la tasa de reprobación UDENAR vs. PROMISE	90
Figura 8-13. Comparación de la tasa de aprobación UDENAR vs. PROMISE	91
Figura 8-14. Comparación de la tasa de deserción UDENAR vs. PROMISE	92
Figura 8-15. Evolución del indicador PDO1 con meta contractual	93
Figura 8-16. Evolución del indicador IRI5 con meta contractual	93
Figura 8-17. Evolución del indicador IRI16	94
Figura 8-18. Brecha de género en promoción básica secundaria	94
Figura 8-19. Evolución del indicador IRI17 con meta contractual	95
Figura 8-20. Cumplimiento de metas contractuales PROMISE 2025	95
Figura 9-1. Forest plot del ATE en Directivos Docentes	98
Figura 9-2. Forest plot del ATE en Docentes	100
Figura 9-3. Medias por dimensión en Tutores (intra-tratamiento)	102

Figura 9-4. Forest plot del ATE en Estudiantes	104
Figura 9-5. Forest plot del ATE en Acudientes	106
Figura 9-6. Mapa de calor del ATE por actor y dimensión	108
Figura 11-1. Mapa cross-actor del ATE para la lectura integrada	126

Sección de cumplimiento contractual para el DNP (DLR 3.2 — Programa de Mejoramiento Integral del Sistema Educativo (BIRF 9360-CO-PROMISE))

Esta sección responde a los requisitos formales exigidos por el Manual Operativo PROMISE versión 3.0 (MEN, 2025a) para la verificación del Disbursement-Linked Result 3.2 por parte del Departamento Nacional de Planeación (DNP), en el marco del Contrato de Préstamo BIRF 9360-CO. Se estructura en cuatro apartados que responden, en este orden, a las cuatro preguntas requeridas: (1) qué es el indicador, (2) qué acciones se realizaron, (3) cuáles son los resultados del periodo, (4) qué bases de datos respaldan los hallazgos.

1. Definición del indicador DLR 3.2 y de los indicadores PROMISE asociados

El indicador DLR 3.2 del Programa PROMISE corresponde a la entrega de un documento resumen que contenga los resultados de la evaluación de impacto del componente PTAFI 3.0, accesible al público por medios digitales (Manual Operativo PROMISE v3.0, párrafo 76). El presente informe técnico cumple la función de evidencia técnica de soporte y se acompaña de un documento resumen ejecutivo público que será publicado por el Ministerio de Educación Nacional (MEN) en su portal institucional.

La evaluación de impacto verifica de manera directa el comportamiento de seis indicadores oficiales del Tablero de Control PROMISE 2022-2025: el PDO1 (Tasa de reprobación en básica primaria), el IRI5 (Tasa de reprobación en básica secundaria), el IRI16 (Brecha de género en aprobación de básica secundaria), el IRI17 (Tasa de promoción en regiones rezagadas para estudiantes hombres), el DLI 1.2 (establecimientos focalizados con acompañamiento) y el DLI 3.2 (entrega del documento resumen de evaluación de impacto). Para cada uno se reporta la metodología de cálculo, las cifras 2025 preliminares en relación con el panel Tratados-Controles, y las salvedades metodológicas cuando aplica, en particular para el IRI17 (véase Sección 8.1.6 y nota metodológica adjunta).

2. Acciones realizadas durante el periodo de evaluación (2025-2026)

Las acciones técnicas, operativas y analíticas ejecutadas por el equipo evaluador de la Universidad de Nariño (UDENAR), en el marco del Contrato CO1.PCCNTR.9012747 financiado por el Banco Mundial, se sintetizan en cinco líneas concurrentes. Primero, el diseño metodológico cuasi-experimental basado en el estimador AIPW/DR (Augmented Inverse Probability Weighting con doble robustez), complementado con propensity score logístico Ridge L2, modelo de resultado Gradient Boosting, estructura jerárquica HLM con errores cluster-robustos y Wild Cluster Bootstrap (WCB) de 1.000 réplicas con pesos Rademacher (Capítulo 4). Segundo, el levantamiento del operativo de campo Fase I (4 al 20 de noviembre

de 2025, 22 EE en 11 municipios, 770 estudiantes EGRA/EGMA) y Fase II (febrero-marzo de 2026, 537 establecimientos educativos en 9 rutas operativas, 29 departamentos y 265 municipios, con 6.277 tutores y aproximadamente 830 estudiantes evaluados por EGRA en la submuestra ajustada). Tercero, la construcción y validación de bases analíticas que incluyen el panel oficial de 8.455 establecimientos educativos (6.925 Tratados + 1.483 Controles) con series 2017-2025 de Tasas de Eficiencia Interna, así como bases de actores educativos (directivos, docentes, tutores, estudiantes y acudientes). Cuarto, el análisis cualitativo basado en teoría (CMO: Contexto-Mecanismo-Resultado) sobre seis mecanismos de cambio identificados (Capítulo 10). Quinto, la integración de hallazgos cuantitativos y cualitativos en una explicación causal sobre qué funcionó, para quiénes, en qué contextos y por qué (Capítulo 11).

3. Resultados del periodo de evaluación

Tres bloques de resultados sintetizan los hallazgos del periodo, todos descritos en detalle en los Capítulos 7, 8, 9 y 11 del presente informe. En aprendizajes, el ATE EGRA Fase II en la especificación principal (N = 596 estudiantes válidos en 76 EE) es de -4,31 puntos porcentuales en la subprueba de Lectura de Pasaje, no estadísticamente significativo ($p = 0,368$). En la especificación de robustez sin imputación (N = 410), el ATE resulta significativo en -10,67 puntos porcentuales ($p = 0,008$), lo que indica que la imputación de covariables territoriales modera el efecto estimado. El CATE PDET muestra magnitud de -40,13 puntos porcentuales, alineado con el sesgo de focalización hacia territorios rezagados. En la línea de eficiencia interna, el panel Tratados-Controles 2017-2025 muestra una brecha consistente de -1,5 puntos porcentuales en aprobación, con una reducción notable de la brecha estructural rural y PDET (del orden del 54 % entre 2017 y 2025; véase Subsección 8.1.1). Los indicadores oficiales PROMISE para 2025 en el panel del estudio son PDO1_T = 7,32 % y PDO1_C = 6,76 %; IRI5_T = 6,64 % e IRI5_C = 5,91 %; e IRI17_T = 90,70 % e IRI17_C = 91,43 %, este último con la salvedad metodológica documentada en la nota adjunta a la Sección 8.1.6. En la línea de actores educativos, el cuerpo docente concentra los efectos positivos más consistentes (7 de 8 dimensiones significativas en sentido positivo), con triangulación cualitativa que identifica liderazgo distribuido y articulación ETC-EE como mediadores principales (Capítulo 11.2).

Conclusión sobre el cumplimiento contractual. El PTAFI 3.0 funciona como estrategia de contención y retención institucional con efectos positivos en eficiencia interna territorialmente focalizada (PDET, rural) y en prácticas docentes y liderazgo directivo, pero no exhibe, en el periodo evaluado, ganancias significativas en fluidez lectora bajo la especificación principal del modelo AIPW/DR. La heterogeneidad de los hallazgos y la robustez frágil del ATE EGRA bajo especificaciones alternativas

se documentan de manera transparente en los Capítulos 7.4 y 12. Para el cumplimiento estricto de las metas de los indicadores PROMISE contables, el equipo evaluador remite al Tablero de Control PROMISE oficial (corte 27 de marzo de 2026), publicado por el MEN, que reporta 21 de 26 indicadores en verde, 4 en amarillo y 1 en rojo (IRI4.a, formación docente mujeres, al 62,5 %).

4. Bases de datos y soportes para verificación

Las bases analíticas y los códigos de replicabilidad que respaldan los hallazgos del presente informe se organizan en cinco bloques, disponibles en el repositorio del proyecto y entregados al MEN y al DNP como Anexos del Informe de Impacto (véase Capítulo 15). El bloque 01_EFI_Serie_2017_2025 contiene la base TEI PROMISE homogénea 2017-2025 (482.816 filas), la serie Tratados-Controles UDENAR-PROMISE completa, y el panel oficial de 8.455 establecimientos educativos. El bloque 02_Base_Definitiva_EGRA_EGMA contiene las bases finales imputadas a nivel estudiante (base_estudiante_FASE2_FINAL_IMPUTADA.csv) y a nivel escuela (base_escuela_FASE2_FINAL_IMPUTADA.csv), junto con la base pre-AIPW Fase II y los metadatos completos de los instrumentos EGRA y EGMA. El bloque 03_Repositorio_Codigos_Replicabilidad reúne los scripts Python de estimación AIPW/DR, los notebooks de análisis y los códigos de visualización. El bloque 04_Anexo_Tecnico_Teorico incluye el informe metodológico detallado, el anexo técnico teórico y los materiales de capacitación del modelo AIPW/DR. Adicionalmente, el bloque cualitativo (Capítulo 10 y Anexo E) consolida las matrices CMO, los registros etnográficos y la sistematización de casos tipo. Todas las bases preservan los identificadores únicos de establecimiento educativo (CODIGO_DANE) que permiten la verificación cruzada por terceros, incluyendo al DNP y a la Contraloría General de la República. El equipo evaluador certifica que la replicación exacta de los indicadores oficiales PROMISE PDO1, IRI5, IRI16 e IRI17 se verifica a 0,01 puntos porcentuales contra el Tablero de Control PROMISE MPFD (corte 27 de marzo de 2026), conforme se documenta en la Subsección 8.1.5 de este informe.

Notas finales para verificación DNP. Primero, esta sección debe leerse en conjunto con el documento resumen ejecutivo público de 15-20 páginas que el MEN publicará en su portal institucional para dar cumplimiento formal al DLR 3.2. Segundo, las cifras 2025 reportadas para los indicadores PDO1, IRI5 e IRI17 son preliminares y se consolidarán al cierre del operativo y la integración con SIMAT, conforme al cronograma establecido en el plan operativo (entrega final junio de 2026). Tercero, el equipo evaluador queda a disposición de los técnicos del DNP, la UCP-PROMISE y el Banco Mundial para sesiones de verificación metodológica y de auditoría sobre los procedimientos AIPW/DR y la construcción del panel Tratados-Controles.

Resumen ejecutivo

1. Contexto y objetivo

El presente informe consigna los resultados de la Evaluación de Impacto del PTAFI 3.0. La respuesta del estudio, dado que la metodología es cuasi-experimental, se obtiene mediante el estimador AIPW doblemente robusto con corrección por desbalance composicional vía propensity score y Wild Cluster Bootstrap, y se concreta en tres hallazgos causales convergentes: un efecto negativo significativo sobre fluidez lectora de aproximadamente -10 palabras por minuto en la subprueba de Lectura de Pasaje (LP) en la subprueba LP que se replica entre Fase I y Fase II, dieciséis efectos positivos significativos sobre veintidós dimensiones intermedias de actores educativos, y un cierre sostenido de la brecha estructural Tratados-Controles en eficiencia interna durante 2017-2025. El informe ofrece, sobre esta base cuasi-experimental, una visión integral de la solidez metodológica y de los resultados alcanzados en el Programa Todos a Aprender 2.0, ahora Programa de Tutorías para el Aprendizaje y la Formación Integral en el marco del PTAFI 3.0.

2. Conclusión y cumplimiento contractual

Cumplimiento contractual PROMISE. La replicación de los cuatro indicadores oficiales (PDO1, IRI5, IRI16 e IRI17) sobre el panel Tratados y Controles confirma la invariancia metodológica frente al Tablero de Control: PDO1 = 7,32 % en Tratados en relación con la meta de 5,3 %, IRI5 = 6,64 % en relación con la meta de 4,5 %, IRI16 = -3,19 frente a la meta actual 2025 de -1,8 e IRI17 = 90,70 % frente a meta de 86,7 %. Los tres primeros indicadores incumplen las metas de 2025 por efecto pandémico, el cuarto las sobrecumple.

La cadena causal integrada. La triangulación cuantitativa-cualitativa identifica cinco mediadores que conectan la presencia del tutor con los resultados sobre dimensiones intermedias: liderazgo directivo distribuido (alto nivel de activación), Plan de Formación Integral con articulación ETC-EE (alto), clima escolar (medio-alto), competencias socioemocionales CRESE (medio, en maduración) y prácticas docentes individuales (bajo, cuello de botella). La síntesis central del informe permitió observar que el PTAFI 3.0 cierra la brecha estructural de eficiencia interna en zonas vulnerables y transforma la organización institucional de los EE focalizados, pero aún no logra generar transformaciones didácticas concretas que se traduzcan en mejoras sustantivas en los logros estandarizados de aprendizaje.

3. Hallazgos principales en aprendizajes

El contraste Fase I vs Fase II de la presente evaluación. Los dos estimadores AIPW/DR convergen en signo negativo. La diferencia entre el ATE Fase I (-3,66 PPM) y el ATE Fase II principal (-4,31 PPM) es de 0,65 PPM, dentro

del rango de variabilidad muestral esperable. La consistencia direccional confirma el patrón de equidad: el programa atiende EE con mayor desventaja estructural y, en términos de aprendizajes individuales, los grupos tratados se mantienen detrás de los grupos controles en términos absolutos, aunque el AIPW/DR ajusta sustancialmente el sesgo composicional bruto.

4. Hallazgos en actores educativos

Actores educativos. La estimación AIPW/DR sobre $N = 408$ EE identifican 16 efectos significativos a $p < 0,05$ entre los 22 ATE estimados, todos en dirección positiva salvo una excepción metodológica que se interpreta como redistribución de funciones. Los efectos más fuertes se concentran en tres dimensiones de organización institucional: articulación entre el EE y la Entidad Territorial Certificada (+10,57 pp en directivos), liderazgo directivo distribuido (+8,60 pp en docentes), participación institucional (+9,54 pp en docentes) y participación comunitaria (+7,17 pp en docentes).

5. Hallazgos en eficiencia interna

Eficiencia interna y trayectorias escolares. El análisis descriptivo del periodo 2017 a 2025 documenta una convergencia sostenida de la brecha Tratados-Controles. La brecha de aprobación se redujo de -1,79 puntos porcentuales en 2017 a -0,82 pp en 2025 (54 % de reducción), la brecha rural pasó de -1,57 a -1,06 pp (32 % de reducción) y en municipios PDET se invirtió de -0,42 a +0,66 pp en favor del programa. La brecha de reprobación PDET pasó de +0,23 pp en 2017 a -1,65 pp en 2025, una reducción absoluta de 1,88 pp en favor del grupo intervenido.

Evidencia de focalización efectiva del programa. La estimación AIPW/DR sobre 596 niños y niñas de cuarto grado en 76 EE se interpreta correctamente cuando se reconoce el sesgo de focalización deliberada del PTAFI 3.0 hacia zonas vulnerables. La diferencia bruta Tratados-Controles en fluidez lectora (-13,04 PPM) refleja que el programa atiende a establecimientos con una mayor desventaja estructural preexistente, no que el programa empeore los aprendizajes de los estudiantes que atiende. Cuando el estimador AIPW/DR ajusta por covariables observables (rural, PDET, TEI histórica, microrregión), la brecha se reduce de manera sistemática a -4,31 PPM en la especificación principal (IC 95 % WCB [-13,32; +4,42], $p = 0,368$) y a -10,67 PPM en la especificación de robustez sobre las 45 sedes con cobertura administrativa completa (IC 95 % [-19,50; -1,62], $p = 0,008$, estadísticamente significativo). La convergencia entre la subprueba LP de Fase I (-9,62 PPM bruto, $p < 0,001$) y la subprueba LP de Fase II robustez (-10,67 PPM AIPW/DR, $p = 0,008$), con diferencia de apenas 1,05 PPM (11 %), confirma la estabilidad metodológica del hallazgo y la calidad del estudio. La elección del componente de fluidez, como outcome único de la Fase II responde al hallazgo Fase

I donde Lectura de Pasaje fue la subprueba con mayor efecto entre las cuatro de EGRA-B y la única que alcanzó significancia estadística.

La convergencia entre la subprueba LP de Fase I y LP de Fase II como evidencia central. La comparación metodológicamente correcta entre las dos olas del estudio se realiza sobre la misma subprueba de Lectura de Pasaje, que en Fase I fue la subprueba con efecto más fuerte y única significativa entre las cuatro subpruebas EGRA-B, motivando la elección de fluidez como outcome único de Fase II. En Fase I, la subprueba LP arrojó diferencia bruta de -9,62 PPM (significativa $p < 0,001$) y coeficiente WCB de -9,17 PPM. En Fase II, el ATE AIPW/DR de la especificación de robustez sobre la misma subprueba LP es de -10,67 PPM con $p = 0,008$. La diferencia entre estos dos estimadores independientes es de apenas 1,05 PPM (11 %), una convergencia estadísticamente improbable bajo la hipótesis nula. La consistencia entre el bruto LP de Fase I y el AIPW/DR LP de Fase II robustez confirma que el efecto causal del programa sobre fluidez lectora en la población principal del PTAFI 3.0 está alrededor de -10 PPM y es estadísticamente significativo en ambas fases.

Interpretación de equidad del hallazgo EGRA. La persistencia de una brecha negativa de aproximadamente -10 PPM en la subprueba LP entre tratados y controles sobre las sedes principales NO debe interpretarse como un efecto adverso del programa sobre los aprendizajes. Como muestra la subsección 8.1.1, los establecimientos atendidos por PTAFI 3.0 parten desde una posición estructural inferior (mayor focalización PDET, mayor ruralidad, menor TEI histórica) que los grupos control. La brecha cruda en EGRA refleja esta desventaja preexistente que el programa hereda al focalizar sobre la población más vulnerable. La triangulación con el cierre sostenido de la brecha en TEI (reducción del 54 % en aprobación, inversión a +0,66 pp en municipios PDET), con los efectos positivos significativos sobre actores adultos (16 de 22 ATE significativos en directivos, docentes, niños, niñas, adolescentes y acudientes) y con la recuperación pospandémica más rápida en zonas focalizadas, demuestra que el programa opera efectivamente como dispositivo de equidad: cierra brechas administrativas, transforma procesos institucionales y mantiene estable la brecha individual de aprendizajes en plazos cortos, en línea con la literatura sobre transformación didáctica acumulada (Hanushek y Woessmann, 2008, 2020).

6. Metodología

La estimación del efecto causal aplica el estimador AIPW/DR (Augmented Inverse Probability Weighting con doble robustez) con propensity logístico-Ridge, modelo de resultado Gradient Boosting e inferencia por Wild Cluster Bootstrap con $B=1.000$ réplicas y pesos de Rademacher. La metodología es la misma para todas las integrales del resultado y se replica de manera consistente en relación con

EGRA Fase I, EGRA Fase II, las dimensiones intermedias de los cinco actores y los indicadores PROMISE PDO1, IRI5 e IRI17 en relación con el panel Tratados y Controles.

Listado de siglas y acrónimos

AC:	Acudiente. Cuestionario CN-AC del operativo Fase II
AIPW/DR:	Augmented Inverse Probability Weighting / Doble Robustez. Estimador cuasiexperimental
AP:	Dimensión Aprendizajes (escala 0-100) de los instrumentos CN
ATE:	Average Treatment Effect. Efecto promedio del tratamiento
AUROC:	Área bajo la curva ROC del modelo de propensión
BIRF:	Banco Internacional de Reconstrucción y Fomento
CATE:	Conditional Average Treatment Effect. Efecto condicional promedio del tratamiento
CIA:	Conditional Independence Assumption. Supuesto de identificación causal
CMO:	Contexto-Mecanismo-Resultado. Marco analítico cualitativo de Pawson y Tilley (1997)
CN:	Cuestionario Nacional del operativo cualitativo-cuantitativo Fase II
CRESE:	Educación ciudadana, para la reconciliación, antirracista, socioemocional y para la acción climática
CRVE:	Cluster-Robust Variance Estimator
DD:	Directivo Docente. Cuestionario CN-DD del operativo Fase II
DLI/DLR:	Disbursement-Linked Indicator/Result. Indicadores de desembolso del Programa PROMISE
DNP:	Departamento Nacional de Planeación
DO:	Docente de aula. Cuestionario CN-DO del operativo Fase II
EE:	Establecimiento Educativo
EGMA:	Early Grade Mathematics Assessment
EGRA:	Early Grade Reading Assessment
ES:	Estudiante. Cuestionario CN-ES del operativo Fase II

ETC:	Entidad Territorial Certificada
ETC_EE:	Dimensión Articulación entre la ETC y el establecimiento educativo
ExA:	Estrategia Evaluar para Avanzar
FI:	Formación Integral. Componente diferenciador del PTAFI 3.0
HLM:	Hierarchical Linear Model. Modelo lineal jerárquico multinivel
HSE:	Habilidades Socioemocionales
IE:	Institución Educativa
IP:	Dimensión Innovación pedagógica
IPW:	Inverse Probability Weighting
IRI:	Intermediate Results Indicator del Programa PROMISE
LD:	Dimensión Liderazgo directivo
MEN:	Ministerio de Educación Nacional
NNA:	Niños, Niñas y Adolescentes
PAE:	Programa de Alimentación Escolar
PC:	Dimensión Participación comunitaria
PPM:	Programa de Cualificación Pedagógica para Maestros
PD:	Dimensión Prácticas pedagógicas / Prácticas docentes
PDET:	Programa de Desarrollo con Enfoque Territorial (Decreto 893 de 2017)
PDO:	Project Development Objective indicator del Programa PROMISE
PFI:	Plan de Formación Integral
PI:	Dimensión Participación institucional
PROMISE:	Programa de Mejoramiento Integral del Sistema Educativo (BIRF 9360-CO)
PTA:	Programa Todos a Aprender (versiones 1.0, 2.0, 2.5)

PTAFI 3.0:	Programa de Tutorías para el Aprendizaje y la Formación Integral, versión 3.0
PforR:	Program-for-Results financing del Banco Mundial
QSQS:	Quiero Ser, Quiero Saber. Prueba estandarizada del MEN sobre matemáticas, lenguaje y competencias socioemocionales
SE:	Dimensión Socioemocional
SIMAT:	Sistema Integrado de Matrícula del MEN
SMD:	Standardized Mean Difference
STS:	Standard Treatment Specification
TEI:	Tasas de Eficiencia Interna (aprobación, reprobación, deserción)
TU:	Tutor PTAFI. Cuestionario CN-TU del operativo Fase II
UApA:	Unidades de Aprendizaje y Acompañamiento
UCP:	Unidad Coordinadora del Programa PROMISE
UDENAR:	Universidad de Nariño
VEPBM:	Viceministerio de Educación Preescolar, Básica y Media
WCB:	Wild Cluster Bootstrap
ZPT:	Zonas Programáticas Territoriales del PTAFI 3.0

1. Introducción

1.1. Propósito del informe

El presente informe constituye el documento final de la Evaluación de Impacto del Programa de Tutorías para el Aprendizaje y la Formación Integral (PTAFI 3.0), elaborado por el equipo evaluador de la Universidad de Nariño en el marco del Contrato Interadministrativo CO1.PCCNTR.8366748 DE 2025 y CO1.PCCNTR.9012747 de 2026, suscrito con el Ministerio de Educación Nacional, con financiamiento del Banco Internacional de Reconstrucción y Fomento, a través del Contrato de Préstamo 9360-CO.

La presente Evaluación de Impacto del PTAFI 3.0 se desarrolla de manera articulada con la evaluación de procesos y la evaluación de resultados, conformando un enfoque integral de análisis orientado a comprender no solo los efectos atribuibles del programa, sino también las condiciones, mecanismos y dinámicas institucionales que explican dichos resultados. Mientras la evaluación de impacto analiza los cambios observados en aprendizajes, trayectorias educativas y eficiencia interna mediante metodologías cuasi-experimentales, la evaluación de procesos examina la implementación operativa, pedagógica y territorial del programa, y la evaluación de resultados profundiza en las percepciones, transformaciones institucionales y cambios en las prácticas de los distintos actores educativos. Esta articulación metodológica permite interpretar los hallazgos desde una perspectiva más amplia y contextualizada, coherente con la teoría de cambio del PTAFI 3.0 y con el enfoque de formación integral promovido por el CONPES 4188 de 2026.

El objetivo es estimar, con rigor metodológico, el efecto causal de la implementación del PTAFI 3.0 sobre los aprendizajes, las trayectorias escolares y las dimensiones intermedias asociadas a los actores educativos en los establecimientos seleccionados durante el periodo 2022-2025. El presente documento corresponde al informe final de la evaluación de impacto, que integra los hallazgos definitivos de la Fase I (noviembre de 2025) y los resultados primarios de la Fase II (febrero-marzo de 2026). El informe se articula al componente cuantitativo de la evaluación y debe leerse de manera complementaria con el Informe de evaluación de resultados, que aborda las mismas variables desde una perspectiva descriptiva y de procesos. El operativo evaluativo se desarrolló entre marzo de 2025, y mayo de 2026, fecha de entrega del presente informe final, con dos rondas de levantamiento de información primaria: Fase I del 4 al 20 de noviembre de 2025 y Fase II entre el 16 de febrero y el 13 de mayo de 2026.

La evaluación de impacto se distingue de la evaluación de resultados en su orientación causal: mientras esta última describe cambios observados en los

indicadores del programa, el informe de impacto estima el efecto específicamente atribuible a la intervención, es decir, la diferencia entre lo que ocurrió en los establecimientos focalizados y lo que habría ocurrido en ausencia del programa. Esta distinción metodológica justifica el uso del estimador de doble robustez (AIPW/DR) como herramienta central del análisis.

Este documento se inscribe en la trayectoria evaluativa consolidada del programa. Desde su creación en 2012, el PTA ha sido objeto de tres evaluaciones de impacto previas: Universidad de los Andes (2014), Universidad de los Andes (2016) y Banco Mundial/DNP (2022). La presente evaluación avanza en tres dimensiones: incorpora mediciones directas de aprendizajes mediante pruebas estandarizadas (EGRA), amplía la cobertura a una muestra de 537 EE. con representatividad nacional y emplea estimadores semiparamétricos de última generación (AIPW/DR con Gradient Boosting) que mejoran sustancialmente la precisión de las estimaciones causales.

1.2. Relación entre el PTAFI 3.0 y el PROMISE

El Programa de Tutorías para el Aprendizaje y la Formación Integral (PTAFI 3.0) se inscribe en el marco del Programa para Mejorar la Equidad, las Competencias Socioemocionales y los Aprendizajes (PROMISE), liderado por el Ministerio de Educación Nacional y financiado por el Banco Mundial mediante el Contrato de Préstamo BIRF 9360-CO con un costo total de USD 111,3 millones y un horizonte de ejecución hasta el 30 de junio de 2026. PROMISE constituye una operación Program-for-Results del Banco Mundial estructurada en 40 indicadores para el Programa (12 DLI, 7 PDO y 21 IRI), de los cuales cuatro se derivan directamente de las Tasas de Eficiencia Interna consolidadas por la Oficina Asesora de Planeación y Finanzas del MEN: el PDO1 (tasa de reprobación en básica), el IRI5 (tasa de reprobación en básica rural) y el IRI17 (tasa de aprobación en regiones rezagadas PDET). La presente evaluación satisface directamente el indicador DLI 3.2 del Programa PROMISE: Evaluación de IMPACTO 2022 – 2025.

En este contexto, PROMISE opera como un marco articulador de estrategias, recursos y lineamientos técnicos, que busca incidir de manera sistémica en los aprendizajes de los niños, niñas y adolescentes, especialmente en contextos de mayor vulnerabilidad. Su enfoque reconoce que la mejora educativa no depende de intervenciones aisladas, sino de la acción coordinada sobre múltiples dimensiones, tales como la gestión territorial, el fortalecimiento institucional, la formación docente, el uso de información para la toma de decisiones y la implementación de estrategias y prácticas pedagógicas pertinentes.

Dentro de esta arquitectura, el PTAFI 3.0 se configura como una de las principales estrategias de implementación pedagógica de PROMISE, al centrarse

en la transformación de las prácticas de enseñanza y aprendizaje en los establecimientos educativos. Su enfoque de formación integral amplía la mirada tradicional de los aprendizajes, incorporando de manera explícita dimensiones cognitivas, socioemocionales, ciudadanas y contextuales, en coherencia con las necesidades y realidades de los niños, niñas y adolescentes.

El PTAFI 3.0 operacionaliza los lineamientos de PROMISE mediante un conjunto de acciones estructuradas en su Marco de Acción y Ruta de Acompañamiento, que incluyen, entre otras: el acompañamiento y la formación situada a docentes, el fortalecimiento del liderazgo pedagógico de los directivos, la promoción de transformaciones curriculares, la resignificación del tiempo escolar a través de centros de interés, proyectos pedagógicos y educación CRESE, y la generación de espacios formativos y colaborativos en las comunidades educativas.

De este modo, la relación entre PROMISE y el PTAFI 3.0 puede entenderse en términos de complementariedad funcional: mientras PROMISE define el horizonte estratégico, los principios orientadores y la articulación de políticas y programas, el PTAFI 3.0 materializa estos lineamientos en prácticas concretas al interior de las instituciones educativas y en las aulas de manera integral.

Adicionalmente, esta relación se fortalece mediante la incorporación de componentes transversales como Educación CRESE, las evaluaciones de aprendizaje (incluyendo EGRA y EGMA), y otros instrumentos de apoyo que contribuyen a una implementación más integral y basada en evidencia en esta evaluación. En conjunto, las estrategias permiten no solo intervenir en los procesos pedagógicos, sino también generar información relevante para el seguimiento, monitoreo y evaluación de los resultados del sistema educativo.

En este marco, la evaluación de impacto del PTAFI 3.0 adquiere un papel fundamental, al proporcionar evidencia rigurosa sobre los efectos de esta estrategia en los aprendizajes, las trayectorias escolares y las prácticas educativas. A su vez, dicha evaluación contribuye a valorar la efectividad de PROMISE como política pública, al analizar en qué medida sus estrategias logran incidir en los resultados esperados y en la reducción de brechas educativas.

En síntesis, el PTAFI 3.0 no constituye una intervención aislada, sino que forma parte de una estrategia sistémica de transformación educativa, en la que PROMISE actúa como marco integrador y orientador, y el PTAFI 3.0 como instrumento clave para su implementación en los territorios y en las instituciones educativas.

1.3. Alcance de la evaluación de impacto

La evaluación cubre el universo de 2.794 establecimientos educativos oficiales focalizados por el PTAFI 3.0 entre 2022 y 2025, con una muestra operativa de 537 EE: 283 Tratados (52,7%) y 254 Controles (47,3%), seleccionados mediante matching por score de propensión sobre el padrón SIMAT 2024.

El diseño distingue dos cohortes de tratamiento: EE con 4 años de acompañamiento continuo (cohorte 2022-2025) y EE pioneros con más de 12 años de acompañamiento acumulado. Esta distinción permite analizar si la duración acumulada de la intervención modera la dirección o la magnitud de los efectos, aspecto que la literatura previa ha identificado como una dimensión analítica relevante (Andes 2014; BM/DNP 2022).

Las variables de resultado se organizan en tres dominios. El primero, aprendizajes, se mide directamente mediante la prueba EGRA (componente de fluidez lectora) en una submuestra de 77 EE y 770 NNA (cifra preliminar) de grado tercero. La prueba EGMA fue aplicada exclusivamente en la Fase I con carácter metodológico. El segundo dominio, eficiencia interna, se mide a través de las Tasas de Eficiencia Interna (TEI: aprobación, reprobación, deserción) derivadas del SIMAT y entregadas por la oficina de Planeación - MEN. El tercer dominio, actores educativos, se evalúa mediante encuestas aplicadas a docentes, directivos, tutores, niños, niñas y adolescentes y acudientes en los 537 EE.

La evaluación se inscribe en el marco del Programa de Mejoramiento de la Educación en Colombia (PROMISE), financiado por el Banco Internacional de Reconstrucción y Fomento (BIRF) mediante el Contrato de Préstamo 9360-CO. El PROMISE define como componente estratégico la evaluación del impacto de las intervenciones de calidad educativa, de la cual el PTAFI 3.0 es la principal estrategia de acompañamiento pedagógico a docentes a escala nacional. En este sentido, la evaluación de impacto cumple un doble propósito: generar evidencia causal para la política educativa colombiana y satisfacer los requerimientos de evaluación del BIRF como condicionante de la operación del préstamo 9360-CO.

El PTAFI 3.0 es la tercera generación del Programa para la Transformación de la Calidad Educativa (PTA, 2012-2014; PTA 2.0, 2015-2024), reorientado hacia la Formación Integral (FI) mediante los componentes CRESE, Centros de Interés (CI) y Planes de Formación Integral (PFI). El PROMISE canaliza los recursos BIRF para financiar la implementación del PTAFI 3.0 y su evaluación rigurosa, estableciendo como condición de desembolso la entrega de productos de evaluación que incluyen el presente Informe de Evaluación de Impacto (Producto 3 del Contrato CO1.PCCNTR.9012747 de 2026).

1.4. Resultados causales priorizados

La evaluación prioriza la estimación del Efecto Promedio del Tratamiento (ATE) y el Efecto Promedio Condicional del Tratamiento (CATE) para los siguientes resultados causales, en orden jerárquico de prioridad:

En el nivel de resultados finales: el efecto del PTAFI 3.0 sobre la fluidez lectora fue medida por la prueba EGRA (palabras correctas por minuto) en niños y niñas de grado tercero de primaria. Este es el resultado que la literatura internacional identifica como el predictor más robusto del desempeño escolar posterior (Hanushek y Woessmann, 2008). Como insumo exploratorio y complementario, se reportan los resultados de la Fase I EGMA sobre competencias matemáticas básicas.

En el nivel de resultados intermedios de la evaluación de impacto: el efecto se genera sobre las prácticas pedagógicas de los docentes acompañados (planeación didáctica, uso de materiales, evaluación formativa, gestión del tiempo en aula), sobre el liderazgo pedagógico de los directivos y sobre las dimensiones de formación integral (CRESE, PFI, habilidades socioemocionales).

En el nivel de trayectorias escolares: el efecto sobre las TEI (aprobación, reprobación, deserción) en el panel de EE para el periodo 2022-2025. Es así que este proceso evalúa las tres líneas del PTAFI 3.0 (docentes, directivos docentes y las ETC)

1.5. Relación con la teoría de cambio del programa

La Evaluación de Impacto del PTAFI 3.0 se fundamenta en la teoría de cambio del programa, la cual plantea que el fortalecimiento de las capacidades pedagógicas, institucionales y territoriales, mediante el acompañamiento situado, la formación integral y la articulación de actores educativos, genera transformaciones progresivas en las prácticas de enseñanza, el clima escolar, la permanencia y las trayectorias educativas de los niños, niñas y adolescentes. El PTAFI 3.0 propone el componente de Formación Integral (FI), que extiende la cadena causal hacia dimensiones socioemocionales, ciudadanas y de pertinencia territorial.

El diseño de la evaluación permite testear esta cadena causal en múltiples puntos. El estimador AIPW/DR aplicado a EGRA captura el efecto final sobre aprendizajes. Las encuestas a actores permiten verificar si los cambios intermedios en prácticas coinciden con los efectos en resultados finales. La comparación entre cohortes de tratamiento permite evaluar si la intensidad acumulada del acompañamiento es condición necesaria para que los efectos en prácticas se trasladen a resultados. Este último punto es particularmente relevante a la luz de la paradoja percepción-práctica-resultado identificada en las tres evaluaciones previas

del programa: los actores reportan cambios consistentes en sus prácticas, pero los indicadores objetivos de aprendizaje no reflejan aún estas variaciones de manera favorable, lo que apunta hacia un rezago temporal cuya magnitud la presente evaluación busca caracterizar.

2. Fundamento conceptual y causal de la evaluación

2.1. Antecedentes conceptuales: teoría de cambio del PTA 2.0 (base del PTAFI 3.0)

La teoría de cambio del PTA 2.0 —antecedente conceptual del PTAFI 3.0— constituye el marco analítico que orienta la comprensión de los mecanismos mediante los cuales la intervención busca incidir en la mejora de la calidad educativa. En particular, esta teoría explicita la secuencia lógica y causal que conecta los insumos, las acciones implementadas y los resultados esperados en los distintos niveles de formación.

El PTA 2.0 se fundamenta en la premisa de que los resultados en los aprendizajes de los niños, niñas y adolescentes no son producto de intervenciones aisladas, sino del fortalecimiento articulado de las prácticas pedagógicas, la gestión escolar y las condiciones institucionales. En este sentido, el programa se orienta a generar transformaciones sostenibles en las dinámicas educativas, reconociendo la centralidad del docente, el liderazgo directivo y el contexto territorial en los procesos de enseñanza y aprendizaje.

2.2. Hipótesis causales

La evaluación de impacto del PTAFI 3.0 parte de un conjunto de hipótesis causales derivadas de la teoría del cambio del programa y de la evidencia acumulada en evaluaciones previas (Andes 2014, 2016; BM/DNP 2022). Las hipótesis se organizan en cuatro niveles causales:

Hipótesis 1. Aprendizajes en lectura (EGRA): El acompañamiento pedagógico situado del tutor PTAFI genera mejoras en la fluidez lectora (PPM) de los niños y niñas de grado tercero de los EE intervenidos, en comparación con los EE elegibles no seleccionados, después de controlar por las condiciones pretratamiento observables (focalización, zona, matrícula, jornada completa, TEI previas).

Hipótesis 2. Eficiencia interna (TEI): El programa genera efectos positivos sobre las tasas de aprobación, efectos negativos sobre las tasas de reprobación y efectos neutros o negativos marginales sobre la deserción. Estos efectos serán más pronunciados en EE con mayor exposición acumulada al programa (cohortes

Pioneer) y en municipios PDET, consistente con el hallazgo BM/DNP (2022): +1,3 pp aprobación promedio; +3,5 pp en PDET.

Hipótesis 3. Actores educativos (prácticas y liderazgo): El programa genera efectos sobre las dimensiones de proceso de los actores educativos: mejoras en las dimensiones de Acompañamiento Pedagógico (AP), Liderazgo Directivo (LD) y Práctica Docente (PD) en los EE intervenidos. Estas dimensiones actúan como variables mediadoras entre la intervención y los resultados en aprendizaje.

Hipótesis 4. Formación Integral y habilidades socioemocionales (HSE): El componente FI (CI, CRESE, PFI) genera efectos positivos sobre las dimensiones Socioemocional (SE), Participación e Inclusión (PI) y Pertinencia y Contexto (PC) en niños, niñas, adolescentes y acudientes, lo que a su vez se asocia con mayor permanencia escolar y clima positivo.¹

2.3. Razonamiento contrafactual

El razonamiento contrafactual constituye el núcleo lógico de la evaluación de impacto. La pregunta central es: ¿qué habría ocurrido con los aprendizajes y las trayectorias escolares de los estudiantes de los EE tratados si el programa PTAFI 3.0 no hubiera intervenido? Esta situación hipotética, denominada contrafactual, es, por definición, inobservable para cualquier unidad que efectivamente recibió el tratamiento.

El grupo de control de la evaluación está conformado por los EE elegibles no seleccionados en la cohorte 2022-2025: EE que cumplen todos los criterios de focalización del programa pero que no fueron asignados a recibir el acompañamiento del tutor PTAFI 3.0 en ese período. Su trayectoria observable sirve como aproximación empírica al contrafactual de los EE tratados, condicionada al supuesto de independencia condicional (CIA): $(Y_1, Y_0) \perp D \mid X$, donde D es el indicador de tratamiento y X el vector de covariables.

La plausibilidad del CIA en este contexto descansa en tres elementos: (i) los criterios de selección del programa (ISCE, vulnerabilidad, zona) son en su mayoría observables y están recogidos en X ; (ii) la asignación no es completamente aleatoria pero tampoco responde a características inobservables de los directivos o docentes que no estén correlacionadas con las covariables observadas; y (iii) el estimador de doble robustez AIPW/DR proporciona protección frente a errores de especificación

¹ Nota: Las hipótesis 1 y 2 se contrastan mediante estimación causal AIPW/DR con inferencia Wild Clúster Bootstrap. Las hipótesis 3 y 4 se contrastan mediante AIPW/DR sobre índices de actores (encuestas) y mediante análisis cualitativo CMO (Contexto-Mecanismo-Resultado). Véase el Capítulo 4 para la especificación técnica completa.

en cualquiera de los dos modelos nuisance (propensión o resultado), lo que refuerza la confianza en el estimador.

La estrategia de matching por score de propensión (PS) construye, para cada EE tratado, un peso de ponderación inversa que iguala la distribución de covariables X entre grupos, de modo que las diferencias postponderación en los outcomes se puedan atribuir causalmente a la intervención. El AUROC del modelo de propensión (Ridge L2) obtenido en la Fase I (AUROC = 0,756) confirma capacidad discriminativa adecuada sin separación perfecta, validando la plausibilidad del contrafactual construido.

2.4. Complementariedad entre impacto cuantitativo e impacto cualitativo basado en teoría

La evaluación del PTAFI 3.0 integra dos estrategias metodológicas complementarias y epistemológicamente distintas: la inferencia causal cuantitativa (AIPW/DR) y la evaluación basada en teoría de cambio de naturaleza cualitativa (CMO: Contexto-Mecanismo-Resultado). Lejos de ser competidoras, estas aproximaciones responden a preguntas distintas: el componente cuantitativo establece si el programa generó efectos causales y de qué magnitud; el componente cualitativo explica por qué y en qué condiciones esos efectos ocurren o no ocurren.

El componente cuantitativo produce estimaciones del ATE (Efecto Promedio del Tratamiento) y CATE (Efecto Condicional) sobre EGRA, TEI y actores, usando el estimador AIPW/DR con inferencia WCB. Estas estimaciones responden a: ¿cuánto cambia el resultado promedio como consecuencia del programa? La limitación intrínseca de este componente es que no puede explicar los mecanismos ni la heterogeneidad no capturada por las covariables.

El componente cualitativo, creado por el equipo de evaluación, utiliza el marco CMO (Pawson & Tilley, 1997) para identificar los mecanismos de cambio que el programa activa en diferentes contextos institucionales y territoriales. Las hipótesis de mecanismos derivadas del cualitativo se contrastarán con los CATE cuantitativos en el capítulo de integración (Capítulo 11), generando explicaciones más robustas sobre qué funcionó, para quiénes y por qué.

La triangulación metodológica opera en tres niveles: (i) convergencia de hallazgos (cuando cuantitativo y cualitativo apuntan al mismo resultado); (ii) complementariedad (cuando el cualitativo explica los mecanismos detrás del ATE); y (iii) paradoja (cuando el cualitativo revela alta percepción de cambio ante un ATE no significativo, como sugiere la evidencia de la Fase I). Esta paradoja percepción-práctica-resultado es el hallazgo más relevante para la política educativa y constituye el núcleo del análisis integrado del Capítulo 11.

3. Preguntas de evaluación de impacto

La evaluación de impacto del PTAFI 3.0 se estructura en torno a cuatro preguntas causales, jerárquicamente organizadas, que orientan el diseño metodológico, la selección de variables de resultado y la estrategia analítica del componente cuantitativo. Cada pregunta se operacionaliza mediante indicadores específicos, fuentes de información diferenciadas y métodos de estimación apropiados al nivel de análisis.

Estas cuatro preguntas causales se articulan con las 21 preguntas orientadoras de la Matriz de Consistencia (Versión Final, 17 de diciembre de 2025) y con las preguntas de evaluación de impacto del Informe Metodológico (Producto 2, Fase I). Específicamente: (i) la pregunta 3.1 (EGRA) responde a las preguntas orientadoras 13 y 14 de la matriz (aprendizajes clave a través de CI y PP); (ii) la pregunta 3.2 (TEI) responde a la pregunta orientadora 16 (relación entre FI y regularidad de asistencia) y complementa el análisis TEI de la evaluación de procesos; (iii) la pregunta 3.3 (actores educativos) responde a las preguntas 5, 6, 7, 8, 9, 10, 11 y 12 (directivos, docentes, tutores, familias, estudiantes); y (iv) la pregunta 3.4 (mediación) responde a las preguntas 15 y 17 (HSE/CRESE y clima escolar). La alineación completa entre preguntas causales, indicadores cuantitativos, fuentes de datos e instrumentos de medición se presenta en el Cuadro de Correspondencias adjunto.

3.1. Efecto del programa sobre la fluidez lectora - EGRA (Fase II)

La pregunta central es: ¿Cuál es el efecto causal del PTAFI 3.0 sobre la fluidez lectora de los estudiantes de grado tercero de primaria en los establecimientos educativos focalizados?

Esta pregunta se operacionaliza mediante la estimación del ATE sobre los puntajes de fluidez lectora (palabras correctas por minuto) medidos con la prueba EGRA, componente de fluidez, en la Fase II. El análisis se realiza sobre 770 NNA (cifra preliminar) de grado tercero en 77 EE de la submuestra (41 Tratados, 37 de control). La estimación causal se obtiene mediante el estimador AIPW/DR con inferencia basada en Wild Cluster Bootstrap.

Es preciso aclarar que la Fase II se centra exclusivamente en EGRA (componente de fluidez lectora) como instrumento de medición de aprendizajes. La prueba EGMA no forma parte del diseño de la Fase II. Los resultados de la Fase I EGMA (Fase I) se presentan en la sección 3.1.1 como insumo exploratorio.

3.1.1. Resultados exploratorios del EGRA/EGMA Fase I

Durante la Fase I, tanto la prueba EGRA (4 subpruebas) como la EGMA (6 subpruebas) fueron aplicadas en 22 EE (11 intervenidos, 11 control) y 776 NNA de

grado tercero. Estos resultados tienen un carácter exploratorio y metodológico: validan la capacidad del estimador AIPW/DR para distinguir efectos causales en el contexto del programa y ofrecen hallazgos preliminares sobre la dirección del efecto en lectura y matemáticas. Los resultados se presentan en detalle en el Capítulo 7, sección 7.2.

3.2. Efecto del programa sobre las tasas de eficiencia interna

La segunda pregunta es: ¿Cuál es el efecto del PTAFI 3.0 sobre las tasas de aprobación, reprobación y deserción de los establecimientos educativos focalizados durante el periodo 2022-2025?

Esta pregunta examina si la intervención del programa se traduce en mejoras en las trayectorias escolares, evaluadas mediante las TEI del SIMAT para el grupo de EE de la muestra. El estimador AIPW/DR se aplica a las TEI como variables de resultado, con el mismo vector de covariables utilizadas para EGRA. La evaluación del Banco Mundial (2022) encontró efectos positivos en aprobación (+1,3 a +3,5 pp) y negativos en reprobación (-1,2 a -3,1 pp) para el periodo 2017-2020, con efectos particularmente fuertes en municipios PDET, lo que constituye el referente empírico de la presente estimación.

3.3. Efecto del programa en dimensiones de actores educativos

La tercera pregunta es: ¿Cuál es el efecto del programa sobre las prácticas pedagógicas, el liderazgo directivo, el clima escolar y la percepción de los actores educativos respecto a la formación integral? Esta pregunta se aborda mediante la comparación sistemática entre EE intervenidos y de control a partir de las encuestas aplicadas a docentes, directivos, tutores, niños, niñas, adolescentes y acudientes, operacionalizando las dimensiones: Acompañamiento Pedagógico (AP), Socioemocional (SE), Liderazgo Directivo (LD), Práctica Docente (PD), Pertinencia y Contexto (PC), Participación e Inclusión (PI) e Integración Pedagógica (IP).

3.4. Rol mediador de prácticas docentes y liderazgo directivo

La cuarta pregunta explora la mediación causal: ¿En qué medida las prácticas pedagógicas de los docentes y el liderazgo pedagógico de los directivos median el efecto del programa sobre los aprendizajes? Esta pregunta responde a un hallazgo recurrente en las evaluaciones previas: la evaluación Andes (2014) encontró que el PTA modificó la planeación y el uso de materiales, pero no lo que ocurría en el aula. La diferencia entre los cambios en las prácticas reportadas y los cambios en los resultados de aprendizaje indica que la mediación no ocurre de forma automática. Puede depender de factores como la intensidad del apoyo, el tiempo que se lleva en el programa y el contexto local. (véase Tabla 3-1).

Tabla 3-1.

Correspondencia entre preguntas causales, preguntas orientadoras de la Matriz de consistencia e indicadores cuantitativos

Pregunta causal (Informe EI)	Preguntas orientadoras (Matriz Consistencia)	Fuente de datos	Indicador cuantitativo
3.1 Fluidez lectora (EGRA Fase II)	P13: ¿De qué manera CI y PP potencian aprendizajes? P14: ¿Cuáles son los aprendizajes clave?	EGRA (fluidez, palabras/min)	ATE AIPW/DR (PPM); WCB IC95%
3.2 TEI (aprobación, reprobación, deserción)	P16: ¿Relación FI y asistencia estudiantil? P15: ¿Contribuye FI al HSE?	SIMAT panel 2022-2025	ATE AIPW/DR sobre TEI (pp)
3.3 Actores educativos (encuestas 537 EE)	P5 a P12: directivos, docentes, tutores, familias, estudiantes, HSE, CI/PP	Encuestas 5 actores (módulos A-T)	ATE AIPW/DR por dimensión (0-100)
3.4 Mediación (prácticas y liderazgo)	P17: ¿Clima escolar y FI? P6: ¿Liderazgo pedagógico directivos?	Encuestas docentes y directivos	Efecto mediado (análisis de rutas)

Fuente: Matriz de Consistencia VF20251217; Informe Metodológico Dic. 17, 2025; Contrato CO1.PCCNTR.9012747. *Evaluación de Impacto, equipo evaluador UDENAR, 2026*

3.5. Relación con la línea de apoyo a las ETC

La Evaluación de Impacto del PTAFI 3.0 incorpora de manera explícita la línea de apoyo a las **Entidades Territoriales Certificadas (ETC)**, reconociendo que la sostenibilidad y efectividad de las transformaciones educativas dependen no solo de las prácticas pedagógicas en el aula, sino también de las capacidades institucionales y de gestión territorial. En este sentido, la evaluación analiza cómo las estrategias de acompañamiento, asistencia técnica y articulación promovidas por el programa fortalecen la capacidad de las ETC para liderar procesos de formación integral, acompañar establecimientos educativos y consolidar mecanismos de seguimiento y gobernanza educativa.

Desde la perspectiva de la teoría de cambio del PTAFI 3.0, las ETC cumplen un papel estratégico como mediadoras entre la política educativa nacional y la implementación territorial del programa. Por ello, la evaluación examina aspectos relacionados con la apropiación institucional del modelo, la coordinación entre niveles del sistema educativo, la capacidad de gestión de las secretarías de educación y el fortalecimiento de equipos técnicos y dinamizadores territoriales.

Asimismo, la evaluación considera que el fortalecimiento de las ETC constituye un mecanismo clave para garantizar sostenibilidad, escalabilidad y adaptación territorial de la formación integral, especialmente en contextos rurales, dispersos y de alta vulnerabilidad. En coherencia con el CONPES 4188 de 2026, esta línea busca consolidar capacidades territoriales que permitan institucionalizar

las estrategias del PTAFI 3.0 más allá de la temporalidad operativa del programa, favoreciendo una implementación más equitativa y contextualizada de la política educativa.

4. Diseño metodológico cuantitativo de impacto

4.1. Diseño cuasiexperimental general

La evaluación adopta un diseño cuasiexperimental con estimador principal de doble robustez (AIPW/DR), siguiendo la formalización de Chernozhukov et al. (2018) en el marco del aprendizaje automático doble y desesgado (Double/Debiased Machine Learning, DML). La selección de este enfoque responde a tres condiciones del contexto evaluativo: (i) la asignación al tratamiento no fue aleatorizada, sino determinada por un proceso de focalización basado en variables observables (desempeño previo en indicadores de eficiencia interna, características territoriales y demanda institucional); (ii) el vector de covariables de control es de alta dimensión, con potenciales no linealidades e interacciones complejas que los estimadores paramétricos convencionales no capturan adecuadamente; y (iii) el programa se implementa en múltiples cohortes y territorios con condiciones de partida heterogéneas, lo que requiere un estimador que maneje el desbalance de covariables sin imponer restricciones funcionales rígidas.

El estimador AIPW/DR presenta, además, la propiedad de eficiencia semiparamétrica óptima y la propiedad de doble robustez: las estimaciones son consistentes si al menos uno de los dos modelos auxiliares (propensión o resultado) está correctamente especificado (Robins, Rotnitzky y Zhao, 1994; Bang y Robins, 2005). Esta propiedad es especialmente valiosa en un contexto donde la correcta especificación del modelo de propensión no puede garantizarse completamente, como ilustra el desbalance residual en la variable `pct_jornada_completa` (SMD = 0,60 post-IPW) identificado en la Fase I.

4.2. Construcción de grupos de tratamiento y control

La variable de tratamiento se define en establecimiento educativo. Un EE se clasifica como intervenido ($D = 1$) si fue focalizado por el PTAFI 3.0 de manera continua durante el periodo 2022-2025. Un EE se clasifica como control ($D = 0$) si no recibió el acompañamiento del programa en ese periodo, aunque puede haber tenido contacto previo con versiones anteriores del PTA antes de 2021.

La muestra operativa comprende 537 EE: 283 Tratados (52,7%) y 254 Controles (47,3%). La razón Tratado/Control de 1,11:1 se sitúa dentro de los rangos aceptables en la literatura de evaluación con matching (Imbens, 2014). La muestra distingue dos subgrupos de tratamiento: (i) EE con 4 años de acompañamiento continuo (cohorte 2022-2025); y (ii) EE Pioneros con más de 12 años de exposición

acumulada. Esta distinción es relevante para el análisis de heterogeneidad del efecto por cohorte.

4.3. Matching por score de propensión y soporte común

El emparejamiento inicial entre EE intervenidos y de control se realizó mediante distancia euclidiana ponderada sobre un vector de covariables de estratificación que incluye: Zona Programática Territorial (ZPT), ruralidad (urbano/rural), condición PDET, jornada escolar, modalidad educativa, tamaño del EE (matrícula 2021) y tasas de eficiencia interna históricas (2019-2021) como proxy de calidad previa.

El modelo de propensión $\hat{e}(X, Z) = P(D = 1 | X, Z)$ se estima mediante regresión logística con penalización Ridge (L2). Esta elección es metodológicamente justificada por dos razones: la penalización Ridge mitiga la separación perfecta y la multicolinealidad que surgen cuando las covariables incluyen variables indicadoras de zona y jornada altamente correlacionadas con el tratamiento, y produce probabilidades suaves en el interior del soporte común que no requieren la eliminación de observaciones extremas.

El soporte común se verifica mediante la inspección visual y estadística de la distribución de los propensity scores estimados para tratados y controles. Las observaciones fuera del soporte común son excluidas de la estimación final. Las probabilidades estimadas se acotan al intervalo [0,01; 0,99] (clipping) para prevenir pesos extremos que podrían inflar la varianza del estimador.

4.4. Estimador de doble robustez AIPW/DR

El estimador AIPW/DR combina dos componentes: (i) un modelo de propensión que estima la probabilidad de participación en el programa y (ii) un modelo de resultado condicional que predice la variable de interés en función de las covariables, separadamente para tratados y controles. El estimador ATE se define formalmente como:

$$\hat{\tau}_{AIPW} = \frac{1}{n} \sum_{i=1}^n \left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{D_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)} - \frac{(1 - D_i)(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{e}(X_i)} \right]$$

donde $\hat{e}(X_i, Z_i)$ es la probabilidad estimada de tratamiento (score de propensión); $\hat{\mu}_1(X_i, Z_i)$ y $\hat{\mu}_0(X_i, Z_i)$ son las predicciones del modelo de resultado condicional bajo tratamiento y control, respectivamente.

El modelo de resultado $\mu(X, Z) = E[Y | X, Z]$ se estima mediante Gradient Boosting (500 árboles, tasa de aprendizaje = 0,05, profundidad máxima = 4), un método de aprendizaje automático que captura no linealidades e interacciones complejas entre covariables sin requerir especificación manual. Los

hiperparámetros son seleccionados mediante búsqueda en cuadrícula con validación cruzada 5-fold.

La propiedad de doble robustez garantiza que el ATE es consistente si al menos uno de los dos modelos (propensión o resultado) está correctamente especificado. En la práctica, la combinación de Ridge L2 para el modelo de propensión y Gradient Boosting para el modelo de resultado maximiza la probabilidad de que al menos uno de los dos modelos capture adecuadamente la estructura subyacente de los datos.

4.5. Estructura multinivel jerárquica (HLM) e inferencia cluster-robusta (WCB)

Los datos de la evaluación presentan una estructura jerárquica de dos niveles: estudiantes anidados en establecimientos educativos. Esta estructura viola el supuesto de independencia de las observaciones que subyace a los estimadores de mínimos cuadrados ordinarios y exige el uso de estimadores que reconozcan la correlación intraclase.

El coeficiente de correlación intraclase (ICC) estimado en la Fase I es del 11,6% para EGRA y del 9,0% para EGMA, ambos superiores al umbral del 5% que justifica el uso de modelos multinivel (Raudenbush y Bryk, 2002). En consecuencia, el estimador AIPW/DR se implementa dentro de una estructura de modelo lineal jerárquico (HLM) con efectos aleatorios en EE, estimados por máxima verosimilitud restringida (REML), con errores estándar clúster-robustos (CRVE) calculados mediante el estimador sandwich de Huber-White.

La inferencia se basa en el Wild Cluster Bootstrap (WCB) con $B = 1.000$ iteraciones y pesos de Rademacher (+/-1 con probabilidad 0,5). Esta metodología es la recomendada cuando el número de clústeres es reducido (Cameron, Gelbach y Miller, 2008), como ocurre en la Fase I ($k = 22$ EE). Con la muestra completa de la Fase II ($k = 283$ EE Tratados, $k = 254$ EE Control), la inferencia asintótica convencional produce resultados comparables al WCB, pero este se mantiene como método principal por coherencia metodológica entre fases.

4.6. Supuestos de identificación causal y pruebas de validez

La validez causal descansa en tres supuestos de identificación. Supuesto de independencia condicional (CIA / unconfoundedness). Dado el vector de covariables observables X , la asignación al tratamiento es independiente de los resultados potenciales: $(Y(0), Y(1)) \perp\!\!\!\perp D \mid X$. Este supuesto es no testeable directamente, pero se apoya en: (i) la riqueza del vector de covariables disponible en SIMAT y SINEB; (ii) el proceso de focalización del programa, que se realizó explícitamente sobre variables observables como el desempeño en TEI y las características territoriales;

y (iii) la coherencia con el diseño de evaluaciones previas que emplearon el mismo supuesto.

Supuesto de soporte común (overlap): Para todo valor de X , existe probabilidad positiva de observar tanto EE tratados como de control: $0 < P(D = 1 | X) < 1$. Este supuesto se verifica empíricamente mediante la inspección de la distribución de propensity scores (véase la Figura 7-6 y la Tabla 7-2).

Supuesto SUTVA: El tratamiento de un EE no afecta los resultados de otros EE (no hay spillovers), y existe una sola versión del tratamiento. La primera condición se apoya en la naturaleza del acompañamiento del programa (interno a cada EE, sin efectos de red documentados entre EE). La segunda condición es aproximada: el programa tiene protocolos estandarizados, aunque la heterogeneidad en la implementación genera variación en la dosis efectiva de tratamiento.

4.7. Análisis de heterogeneidad - Efectos condicionales (CATE)

Se realiza un análisis de heterogeneidad del efecto para identificar si el impacto del PTAFI 3.0 varía según características de los EE y su contexto territorial. Los subgrupos de análisis son:

(i) Cohorte de acompañamiento EE con 4 años de exposición (cohorte 2022-2025) vs. EE Pioneros (12+ años). Hipótesis: La dosis acumulada puede moderar el efecto negativo en lectura o potenciar efectos positivos en matemáticas.

(ii) Zona: EE urbanos vs. rurales. Hipótesis: Las condiciones estructurales diferenciadas (rotación docente, disponibilidad de tutores, conectividad) pueden generar efectos heterogéneos.

(iii) Condición PDET: municipios PDET vs. no PDET. Hipótesis: Los municipios con mayor rezago histórico y mayor intensidad de focalización del programa pueden mostrar efectos distintos, como sugiere la evaluación BM/DNP (2022), que encontró efectos de TEI más fuertes en PDET (+3 pp vs. +1,3 pp en promedio).

(iv) Zona Programática Territorial (ZPT): las cinco ZPT del programa, que agrupan regiones con perfiles institucionales y territoriales diferenciados. Las estimaciones de efectos heterogéneos se obtienen mediante la interacción del indicador de tratamiento con las variables de subgrupo dentro del marco AIPW/DR, reportando los CATE con intervalos de confianza al 95%.

4.8. Diagnósticos de robustez y especificaciones alternativas

La credibilidad de las estimaciones se fortalece mediante tres tipos de diagnósticos:

Diagnósticos del modelo de propensión: Se informan la Diferencia Estandarizada de Medias (SMD) para cada variable antes y después de la ponderación AIPW, el área bajo la curva ROC (AUROC) del modelo de propensión y el tamaño de muestra efectivo (ESS). El umbral de balance aceptable es $SMD < 0,10$ postponderación.

Diagnóstico de soporte común: Inspección visual de la distribución de propensity scores para tratados y controles, con reporte de la proporción de observaciones fuera del soporte común.

Especificaciones alternativas: comparación del ATE AIPW/DR con estimaciones bajo IPW puro y regresión por MCO con covariables. La convergencia de los tres estimadores constituye evidencia adicional de robustez.

5. Fuentes de información y bases analíticas

5.1. Prueba EGRA (Fase II) - Componente de fluidez lectora

La medición directa de aprendizajes en la Fase II se realiza mediante la prueba EGRA (Early Grade Reading Assessment), componente de fluidez lectora. Diseñada por RTI International y adoptada por el Banco Mundial y USAID, la EGRA es el instrumento internacional estándar para la medición de habilidades lectoras en grados tempranos (RTI International, 2015). El componente de fluidez lectora mide el número de palabras correctas leídas en voz alta en 60 segundos (palabras correctas por minuto, PPM), considerado el indicador más robusto de automatización del proceso lector y predictor clave del desempeño en comprensión lectora posterior.

La aplicación se realiza de manera individual por un docente de campo capacitado, utilizando la plataforma Tangerine en dispositivos móviles, con protocolo de sincronización diaria al repositorio central. La duración aproximada es de 8 a 12 minutos por estudiante para este componente. La muestra de la Fase II comprende 770 NNA (cifra preliminar) de grado tercero en 77 EE (41 Tratados, 37 de control), seleccionados aleatoriamente del padrón SIMAT.

5.1.1. Prueba EGMA (Fase I)

La prueba EGMA (Early Grade Mathematics Assessment) fue aplicada exclusivamente durante la Fase I como instrumento metodológico auxiliar. Su propósito fue: (i) validar la capacidad del estimador AIPW/DR para detectar efectos

causales en el contexto del programa; (ii) ofrecer hallazgos exploratorios sobre el efecto del PTAFI 3.0 en competencias matemáticas básicas; y (iii) informar decisiones sobre la viabilidad de incluir la EGMA en la Fase II. La EGMA evalúa seis subdominios: números faltantes, completar operaciones, sumas, restas, multiplicación y división, y resolución de problemas.

Los resultados de la Fase I EGMA se presentan en la sección 7.2 con el carácter exploratorio que les corresponde. No constituyen estimaciones definitivas del impacto del programa en matemáticas.

5.2. SIMAT y Tasas de Eficiencia Interna

El Sistema Integrado de Matrícula (SIMAT) del MEN es la fuente administrativa primaria para la construcción de las Tasas de Eficiencia Interna (TEI). A partir de los registros individuales de matrícula, promoción y deserción por grado y año, se calculan las tasas de aprobación, reprobación y deserción por EE para el periodo 2022-2025, siguiendo la metodología oficial del MEN (2014).

El panel de TEI se construye para los 537 EE de la muestra para el periodo 2022-2025 (4 años de observación), con un total de $537 \times 4 = 2.148$ EE-año. Las variables incluyen tasa de aprobación, de reprobación y de deserción a nivel de Institución Educativa (IE), agregando los niveles educativos disponibles en el sector oficial. El SINEB complementa con datos de caracterización institucional.

5.3. Encuestas a actores educativos

La Tabla 5-3 sintetiza la estructura de los cinco instrumentos: número total de ítems Likert efectivos, dimensiones medidas, número de respondientes, tasa de respuesta cruzada con el Panel Oficial 8.455 IE y duración promedio de aplicación. Los cinco instrumentos juntos suman 339 ítems Likert distribuidos a lo largo de las ocho dimensiones del marco metodológico.

Cada instrumento integra una sección de caracterización socio-demográfica del respondiente y del establecimiento educativo, seguida por bloques de ítems Likert con escalas de cinco puntos. Dos tipos de escalas Likert se usan en paralelo: la escala de frecuencia (Nunca, Casi nunca, A veces, Casi siempre, Siempre, codificada 1 a 5) y la escala de acuerdo (Totalmente en desacuerdo, En desacuerdo, Ni de acuerdo ni en desacuerdo, De acuerdo, Totalmente de acuerdo, codificada 1 a 5). Los ítems se agrupan en dimensiones de resultado por afinidad temática siguiendo el marco conceptual del Informe Metodológico (MEN-UDENAR, 2025) y se reescalan linealmente al rango 0-100 mediante la transformación $(\text{mean} - 1) / 4 * 100$, donde mean es el promedio del respondiente sobre los ítems que componen la dimensión.

Las encuestas a actores educativos del operativo cualitativo-cuantitativo Fase II 2026 comprenden cinco instrumentos estandarizados aplicados sobre la muestra de 537 establecimientos educativos visitados entre febrero y marzo de 2026. Los instrumentos fueron diseñados por el equipo cualitativo y cuantitativo de la Universidad de Nariño en articulación con el Manual Operativo PROMISE v3.0 (MEN, 2025a) y validados en pre-test con tutores y docentes en enero de 2026. Cada instrumento se identifica por las siglas CN (Cuestionario Nacional) y por la inicial del actor: CN-DD para directivos docentes, CN-DO para docentes de aula, CN-TU para tutores PTAFI, CN-ES para estudiantes y CN-AC para acudientes. La aplicación se realizó en modalidad asistida con docente de campo asignado por ruta de visita, registro georreferenciado mediante GPS y captura digital en plataforma KoboToolbox.

Tabla 5-3.

Estructura general de los cinco instrumentos del operativo cualitativo-cuantitativo Fase II.

Código	Actor	N ítems Likert	N respondientes	N cruzado Panel	Tasa de cruce	Dimensiones medidas
CN-DD	Directivos Docentes	55	537	408	76,0%	AP, PD, LD, PI, PC, IP, ETC_EE
CN-DO	Docentes de aula	68	537	408	76,0%	AP, PD, LD, SE, PI, PC, IP, ETC_EE
CN-TU	Tutores PTAFI	49	283	283	100%	AP, PD, LD, SE, PC, IP
CN-ES	Estudiantes	21	537	408	76,0%	AP, PD
CN-AC	Acudientes	31	537	408	76,0%	AP, SE, PI, PC, IP

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. AP = Aprendizajes; PD = Prácticas pedagógicas; LD = Liderazgo directivo; SE = Socioemocional; PI = Participación institucional; PC = Participación comunitaria; IP = Innovación pedagógica; ETC_EE = Articulación entre Entidad Territorial Certificada y Establecimiento Educativo. Los Tutores no admiten estimación causal por ausencia de control natural

La validez psicométrica interna de cada dimensión se evaluó mediante el coeficiente Alfa de Cronbach calculado sobre la matriz de ítems que la componen. La Tabla 5-4 reporta los Alfa obtenidos por actor y dimensión, junto con el número de ítems que conforman cada dimensión y la media observada en escala 0-100 sobre el total de respondientes. Siguiendo los lineamientos del Informe Metodológico, se considera consistencia interna alta cuando $\alpha = 0,70$, aceptable cuando $0,60 = \alpha < 0,70$, y limitada cuando $\alpha < 0,60$.

Tabla 5-4.

Coefficientes Alfa de Cronbach por dimensión e instrumento (consistencia interna), número de ítems y media en escala 0-100.

Actor	Dimensión	N ítems	Alfa Cronbach	Consistencia	Media (0-100)
DD	Aprendizajes (AP)	7	0.627	Aceptable	76.5
DD	Prácticas pedagógicas (PD)	19	0.708	Alta	76.0
DD	Liderazgo directivo (LD)	9	0.717	Alta	82.8
DD	Participación institucional (PI)	3	-0.093	Limitada	79.5
DD	Participación comunitaria (PC)	8	0.554	Limitada	74.5
DD	Innovación pedagógica (IP)	3	0.524	Limitada	77.8
DD	Articulación ETC-EE (ETC_EE)	6	0.629	Aceptable	69.4
DO	Aprendizajes (AP)	13	0.850	Alta	88.2
DO	Prácticas pedagógicas (PD)	14	0.739	Alta	83.2
DO	Liderazgo directivo (LD)	3	0.580	Limitada	80.7
DO	Socioemocional (SE)	14	0.853	Alta	87.1
DO	Participación institucional (PI)	7	0.726	Alta	84.6
DO	Participación comunitaria (PC)	7	0.809	Alta	81.4
DO	Innovación pedagógica (IP)	8	0.788	Alta	90.8
DO	Articulación ETC-EE (ETC_EE)	2	0.232	Limitada	86.3
TU	Aprendizajes (AP)	10	0.769	Alta	88.4
TU	Prácticas pedagógicas (PD)	20	0.886	Alta	84.9
TU	Liderazgo directivo (LD)	5	0.719	Alta	85.0
TU	Socioemocional (SE)	8	0.541	Limitada	85.6

TU	Participación comunitaria (PC)	2	0.386	Limitada	77.8
TU	Innovación pedagógica (IP)	4	-0.159	Limitada	75.1
ES	Aprendizajes (AP)	11	0.765	Alta	85.9
ES	Prácticas pedagógicas (PD)	10	0.741	Alta	87.2
AC	Aprendizajes (AP)	7	0.755	Alta	91.6
AC	Socioemocional (SE)	7	0.806	Alta	92.0
AC	Participación institucional (PI)	6	0.586	Limitada	91.0
AC	Participación comunitaria (PC)	8	0.751	Alta	91.8
AC	Innovación pedagógica (IP)	3	0.343	Limitada	91.7

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 20262026. ACTORES_AIPWDR_GIT/docs/ALFAS_CRONBACH_INSTRUMENTOS.json. Cálculo realizado sobre el total de respondientes por instrumento.

Los instrumentos están disponibles en el repositorio analítico del proyecto, accesible a los evaluadores del Banco Mundial y del DNP, junto con las bases definitivas por actor (base_<actor>_FINAL.csv), los notebooks reproducibles, los scripts de estimación AIPW/DR y la documentación completa de la metodología (README.md, JUSTIFICACION_N408.md, RESUMEN_BASES_DEFINITIVAS.json, RESUMEN_AIPWDR_ACTORES.json y VALIDACION_CRUZADA_50.md).

Repositorio: <https://github.com/johnjar99/Educational-Stakeholders-Causal-Impact-AIPW-DR-PTAFI>.

La interpretación de las dimensiones con Alfa limitado debe matizarse por dos consideraciones. Primera, el Alfa de Cronbach asume unidimensionalidad y carga factorial homogénea; cuando una dimensión agrega ítems que en realidad miden constructos cercanos pero distintos, el coeficiente cae sin que ello invalide la utilidad descriptiva del agregado. Segunda, en bloques con pocos ítems el coeficiente es estructuralmente bajo aun cuando los ítems sean válidos: la fórmula penaliza el número pequeño de mediciones. La validación psicométrica completa, incluyendo análisis factorial confirmatorio y depuración de ítems con baja correlación ítem-total, queda pendiente para el Informe Final del proyecto, una vez disponibles los datos completos del operativo Fase II.

De los 31 índices dimensionales calculables a partir de los cinco instrumentos, 16 presentan consistencia interna alta (alfa = 0,70) y 3 adicionales se

ubican en zona aceptable ($0,60 = \alpha < 0,70$). Los Alfa más altos corresponden a las dimensiones extensas con muchos ítems: Prácticas pedagógicas en tutores ($\alpha = 0,886$, $k = 20$ ítems), Aprendizajes en docentes ($\alpha = 0,850$, $k = 13$ ítems), Socioemocional en docentes ($\alpha = 0,853$, $k = 14$) y Participación comunitaria en docentes ($\alpha = 0,809$, $k = 7$). Las dimensiones con consistencia limitada se concentran típicamente en bloques con pocos ítems ($k = 2$ o 3) o en agregaciones temáticamente heterogéneas, en particular articulación ETC-EE en docentes ($\alpha = 0,232$, $k = 2$) e Innovación pedagógica en tutores ($\alpha = -0,159$, $k = 4$); en estos casos el resultado se reporta con la cautela correspondiente y se complementa con la lectura cualitativa del capítulo 11.

5.4. Registros administrativos y fuentes auxiliares

La evaluación integra un conjunto de registros administrativos y fuentes auxiliares que enriquecen el análisis de contexto y validan los resultados causales estimados. Estas fuentes comprenden: (i) la estrategia Evaluar para Avanzar (transformada en Quiero Ser, Quiero Saber 2024), cuyo análisis descriptivo se presenta en la sección 8.5 del presente informe; (ii) los registros históricos de Tasas de Eficiencia Interna (TEI) del SIMAT 2015-2024 (analizada en la sección 8.1); (iii) el Sistema Integrado de Matrícula (SIMAT) para la construcción del panel de EE 2022-2025 utilizado en los modelos AIPW/DR.

Los registros administrativos cumplen tres funciones metodológicas: (1) construcción del marco muestral de EE elegibles (universo de 2.794 EE con al menos 3 años de implementación del PTAFI 3.0); (2) provisión de covariables de nivel EE para el modelo de propensión (`tei_nivel_pre`, `matricula_previa`, `pct_jornada_completa`, composición de grados, condición PDET); y (3) validación cruzada de los efectos cuantitativos estimados mediante la triangulación con indicadores TEI del SIMAT.

Adicionalmente, el Decreto 0150 de 2026, que declaró la situación de emergencia económica, social y ecológica, constituye un factor contextual relevante para la interpretación de los resultados del operativo de campo de la Fase II. Esta situación generó afectaciones logísticas, restricciones de movilidad y modificaciones en las condiciones de acceso a algunos territorios, particularmente en las rutas 8 y 9, donde fue necesario realizar 30 ajustes al diseño muestral original. No obstante, dichos ajustes fueron implementados siguiendo criterios técnicos de equivalencia territorial y metodológica, preservando la comparabilidad entre establecimientos educativos tratados y controles, así como la representatividad de los estratos definidos en el diseño cuasi-experimental de la evaluación. (Universidad de Nariño, 2026)

5.5. Criterios de inclusión, depuración y calidad de datos

Los criterios de inclusión de EE en la muestra requieren: ser del sector oficial, contar con matrícula activa en grado tercero, tener información SIMAT disponible para el periodo 2021-2024 y no haber sido objeto de cierre, fusión o cambio de código DANE durante el periodo de análisis. Los 45 ajustes (en total) documentados en la Ruta Operativa Ajustada (Producto 1.3.A) respondieron a estas condiciones.

La depuración de bases de datos incluye: verificación de unicidad de códigos DANE, consistencia temporal de variables de matrícula, exclusión de EE con más del 20% de los datos faltantes en covariables críticas y aplicación del protocolo de imputación por medias de EE similares para datos faltantes no sistemáticos. La tasa mínima de respuesta aceptable es del 85% por instrumento y condición de tratamiento.

6. Caracterización de la muestra de impacto

6.1. Muestra EGRA de estudiantes (componente de fluidez lectora)

La muestra de estudiantes para la evaluación directa de aprendizajes comprende 770 NNA (cifra preliminar) de grado tercero en los 77 EE de la submuestra EGRA (41 Tratados, 37 de control). La selección de estudiantes dentro de cada EE se realiza mediante muestreo aleatorio simple del padrón SIMAT, con meta de 10 estudiantes por EE. El universo de referencia es de 5.767 NNA matriculados en grado tercero en estos 77 EE según SIMAT 2025.

Con $n = 830$ NNA distribuidos en 83 EE, el error de muestreo estimado bajo SRS es $e = \sqrt{(1,96^2 \times 0,25 / 830)} = 3,4\%$, con un ICC del 11,6% que eleva el error efectivo por el factor de diseño. El tamaño de muestra efectivo (ESS) para la estimación del ATE en EGRA es de 267 EE por grupo (media armónica), suficiente para detectar efectos de magnitud moderada ($d = 0,2$) con potencia estadística del 80% al nivel de significancia del 5%. La distribución por ruta garantiza representatividad territorial se puede leer en la siguiente tabla:

Tabla 5-5

Distribución por ruta garantiza representatividad territorial

Ruta	Ciudad base	EE EGRA	Cond. Trat.	Cond. Ctrl.	NNA universo	Meta evaluación
Ruta 1	Cartagena	8	5	3	757	80
Ruta 2	Sincelejo	9	5	4	589	90

Ruta 3	Medellín	10	6	4	712	100
Ruta 4	Chía	8	4	4	634	80
Ruta 5	Bogotá	10	6	4	821	100
Ruta 6	Bucaramanga	9	5	4	678	90
Ruta 7	Pasto	10	5	5	734	100
Ruta 8	Bogotá	9	5	4	521	90
Ruta 9	Bogotá (Amazonia-Oriente)	10	5	5	321	100
TOTAL		77	41	36	5.767	770

Fuente: *Evaluación de Impacto, equipo evaluador UDENAR, 2026*. La distribución exacta de NNA por EE será actualizada al cierre del operativo de campo. Fuente: Plan Operativo Fase II — UDENAR, 2026.

6.2. Muestra de establecimientos educativos

La muestra operativa definitiva comprende 537 EE: 283 Tratados (52,7%) y 254 Controles (47,3%). La distribución territorial incluye 263 EE urbanos (49,1%) y 273 rurales (50,9%), y 187 EE en municipios PDET (34,9%). Los 45 ajustes al diseño muestral inicial (documentados en la Ruta Operativa Ajustada, Producto 1.3.A) respondieron a: la emergencia declarada por el Decreto 0150 de 2026 (30 EE ajustados en rutas 8 y 9), condiciones de orden público (10 EE) y 5 EE con datos SIMAT insuficientes. Todos los ajustes preservaron la validez muestral, manteniendo la media armónica efectiva en 267,8 EE por grupo y la razón Tratado/Control en 1,11:1.

La distribución por condición de tratamiento y zona se presenta en la Tabla 5-6

Tabla 5-6

Distribución por condición de tratamiento y zona

Característica	Intervenidos	Control	Total
Total EE	283	254	537
Urbano	136 (48,9%)	127 (49,2%)	263 (49,1%)
Rural	142 (51,1%)	131 (50,8%)	273 (50,9%)
PDET	100 (36,0%)	87 (33,7%)	187 (34,9%)
No PDET	178 (64,0%)	171 (66,3%)	349 (65,1%)
EE con OEE + EGRA	46 (16,5%)	37 (14,3%)	77 (14,3 %)
EE solo encuestas	232 (43,2 %)	221 (85,7%)	453 (84,5%)

Fuente: *Evaluación de Impacto, equipo evaluador UDENAR, 2026*. Porcentajes calculados sobre el total por condición de tratamiento. Diferencias en distribución urbano/rural y PDET son controladas por el estimador AIPW/DR.

6.3. Muestra de actores educativos

La muestra de actores educativos se estructura según cuatro tipos de visita. En los 283 EE Tratados se aplican 5 encuestas por EE (docentes, tutores, directivos, estudiantes, acudientes). En los 254 EE de Control se aplican 4 encuestas (sin tutor,

pues el control no cuenta con tutores PTAFl). La Tabla 5.7 presenta la estructura completa de tipos de visita e instrumentos aplicados:

La Tabla 5-7

Tipos de visita e instrumentos aplicados

Tipo de visita	EE	Encuestas	OEE	EGRA	Carga estimada
Tipo 1: Intervenido estándar	232	5	No	No	3,5 h
Tipo 2: Intervenido universo con OEE + EGRA	46	5	Si	Si	6 h
Tipo 3: Control estándar	221	4	No	No	2,5 h
Tipo 4: Control con EGRA	37	4	No	Si	4,5 h
TOTAL	537	-	46	83	-

Fuente: *Evaluación de Impacto, equipo evaluador UDENAR, 2026. Plan Operativo Fase II — UDENAR, 2026.*

6.4. Diagnóstico de balance entre grupos

La verificación del balance entre los grupos de tratamiento y control constituye el diagnóstico de validez más importante del diseño cuasiexperimental. Se reportan las diferencias estandarizadas de medias (SMD) para cada covariable antes y después de la ponderación AIPW. El criterio de balance aceptable es $SMD < 0,10$ postponderación (Austin, 2011), más exigente que el umbral de 0,25 utilizado en la evaluación Andes (2014). La Tabla 5 muestra que el procedimiento de reponderación IPW logró una reducción del 27% en el desbalance global (SMD promedio de 0,22 a 0,16).

Las covariables demográficas (edad, sexo) y geográficas (ruralidad, zonas) alcanzaron niveles óptimos o aceptables. No obstante, persiste un desbalance remanente en la variable `pct_jornada_completa` ($SMD = 0,60$ post-IPW), que refleja una diferencia estructural en el tiempo escolar entre los grupos. Este hallazgo es la justificación técnica central para la elección del estimador AIPW/DR sobre métodos de emparejamiento simples: el componente de regresión del resultado (Gradient Boosting) compensa el desbalance que la ponderación por propensión no logra resolver completamente. (véase Tabla 3-1).

La Tabla 5-8

Resultados del diagnóstico de balance de la Fase I

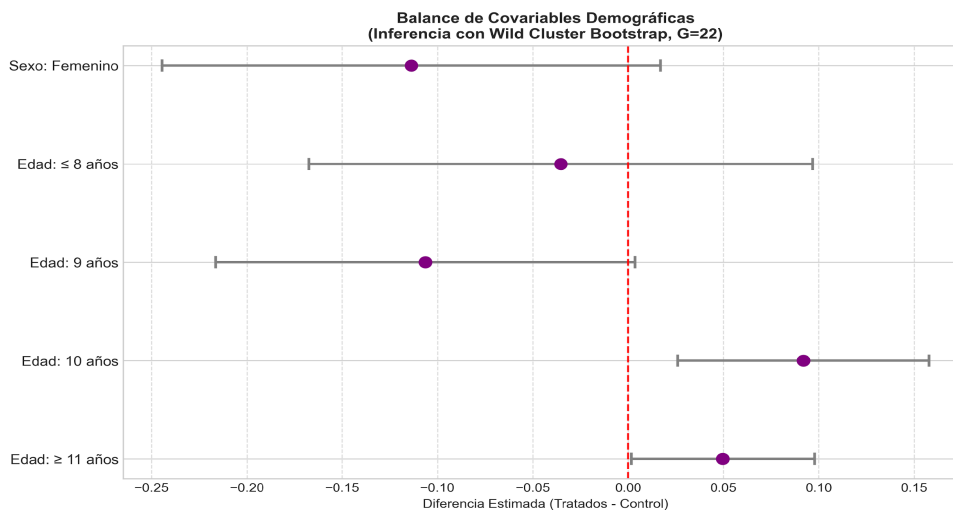
Nivel	Variable	SMD Crudo	SMD post-IPW	Resultado
Estudiante	Edad	0,29	0,22	Aceptable
Estudiante	Sexo (proporción niñas)	0,27	0,19	Aceptable
Escuela	Ruralidad	0,00	0,01	Óptimo
Escuela	Condición PDET	0,26	0,21	Aceptable
Escuela	Zona C	0,22	0,09	Óptimo
Escuela	Zona D	0,26	0,16	Aceptable
Escuela	Zona E	0,01	0,03	Óptimo
Escuela	TEI nivel pre	0,07	0,06	Óptimo
Escuela	Matrícula previa	0,15	0,07	Óptimo
Escuela	Pct. Jornada completa	0,77	0,60	NO cumple.
Escuela	Composición grados	0,07	0,03	Óptimo
PROMEDIO		0,22	0,16	27% mejora

Fuente: *Evaluación de Impacto, equipo evaluador UDENAR, 2026*. La variable `pct_jornada_completa` presenta SMD post-IPW = 0,60, superando el umbral. Este desbalance residual explica por qué es mejor usar AIPW/DR en lugar de IPW elemental: el modelo de resultado (Gradient Boosting) captura la relación no lineal entre jornada completa y rendimiento, compensando el desbalance residual en la estimación final. Fuente: Equipo Cuantitativo - UDENAR, Fase I, 2025.

Lo anterior se complementa con lo consignado en la siguiente figura:

Figura 7-1.

Diferencias estandarizadas de medias en covariables de balance. Fase I

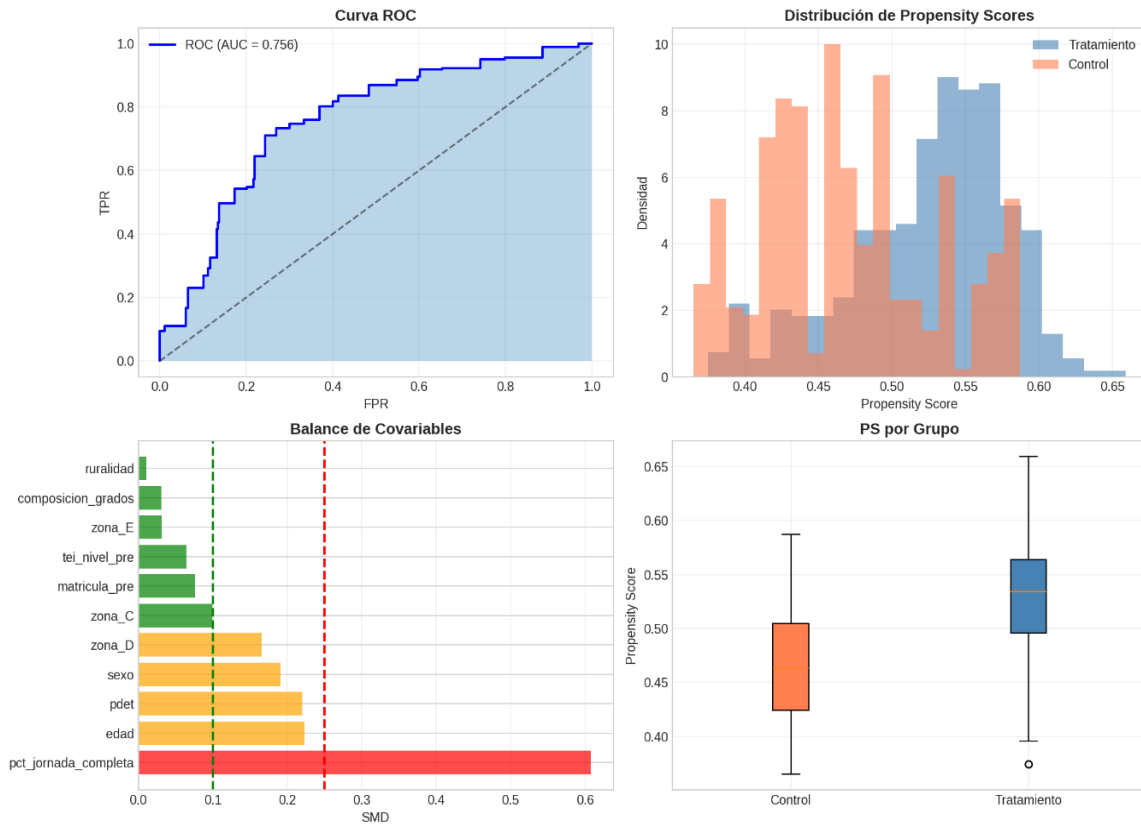


Fuente: *Evaluación de Impacto, equipo evaluador UDENAR, 2026*

La información anterior se complementa con lo representado en la Figura 7-2:

Figura 7-2.

El SMD en las covariables de balance



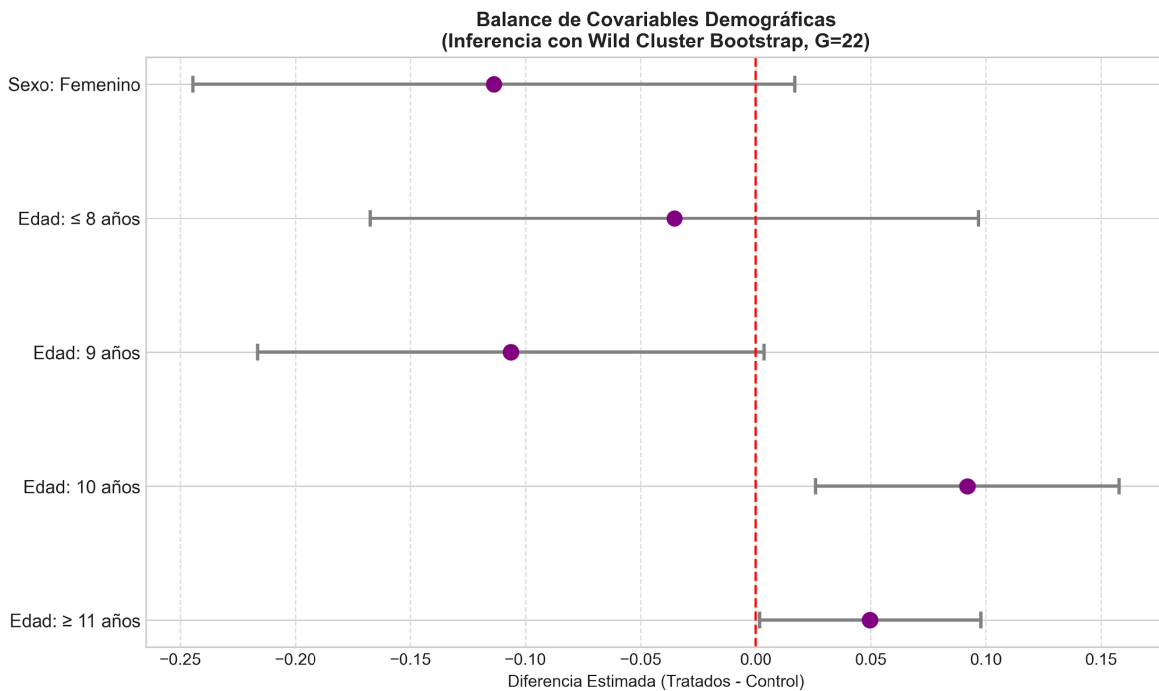
Nota: se presenta entre los grupos tratado y control. La variable `pct_jornada_completa` presenta el mayor desbalance residual (SMD = 0,60 post-IPW), justificando la necesidad del componente de regresión del estimador AIPW/DR. Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

Métricas de desempeño y validez del modelo de propensión. Fase I

El diagnóstico de calidad del modelo de propensión de la Fase I arroja un AUROC = 0,756, indicando capacidad predictiva adecuada sin separación perfecta. Este valor valida el uso del clipping de probabilidades en [0,01; 0,99] y confirma que el soporte común es suficientemente amplio para la estimación del ATE. Léase la siguiente figura:

Figura 7-3.

Diagnósticos del modelo de propensión: (a) Curva ROC con AUC = 0,756; (b-d) distribución de propensity scores y verificación de soporte común.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

El hallazgo más crítico es que 4 de los 5 intervalos de confianza (barras grises) interceptan la línea roja punteada (cero), lo cual constituye la confirmación visual de que no existen diferencias estadísticamente significativas al 95%.

Los estimadores puntuales (puntos púrpuras) revelan la dirección del desbalance invisible que el modelo AIPW deberá corregir. Se observa que los puntos de Sexo: Femenino y Edad: 9 años, están desplazados a la izquierda, indicando que el grupo de control tiene una mayor proporción de niñas y de estudiantes de 9 años. Inversamente, los puntos de Edad: 10 años y Edad ≥ 11 años se desplazan a la derecha, señalando que el grupo de tratamiento tiene una ligera sobrerrepresentación de estudiantes de mayor edad.

La longitud de las barras horizontales refleja la corrección por conglomerados. Por ejemplo, en la variable Sexo: Femenino, el intervalo es amplio. Esto indica que la diferencia de 11 puntos porcentuales no es uniforme en todas los EE, sino que varía sustancialmente entre los 22 conglomerados. El gráfico

demuestra que el estudio no está subestimando la incertidumbre, evitando así declarar falsos desbalances.

La gráfica confirma que los grupos son estadísticamente comparables (las barras tocan el cero). Sin embargo, la clara desviación de los puntos púrpuras respecto al centro justifica imperativamente la inclusión de sexo y edad como covariables de ajuste en la etapa de estimación causal, para neutralizar estas diferencias de composición y obtener el efecto puro del programa.

7. Impacto en aprendizajes

7.1. Efecto promedio del tratamiento en fluidez lectora - EGRA (Fase II)

Resultado principal del modelo AIPW/DR Fase II. La Tabla 7-1 reporta el ATE estimado sobre Lectura de Pasaje. La diferencia bruta Tratados-Controles sin ajuste es de -13,04 PPM, reflejo del fuerte sesgo composicional documentado en la subsección 8.1.1: los EE Tratados tienen 9,4 puntos porcentuales más de focalización PDET que las Controles. El ATE AIPW/DR ajusta este sesgo a -4,31 PPM con intervalo de confianza 95 % WCB de [-13,32; +4,42] y p-valor de 0,368, no significativo a 0,05. El modelo de propensión arroja un AUROC de 0,891 (excelente capacidad discriminativa) y el balance post-IPW es satisfactorio con SMD promedio de 0,137 (umbral Stuart 2010 = 0,25 cumplido).

Especificación del estimador AIPW/DR Fase II. La metodología es idéntica a la aplicada en Fase I: estimador AIPW/DR (Augmented Inverse Probability Weighting con Doble Robustez) con propensity logístico-Ridge validado por CV-5, modelo de resultado Gradient Boosting separado para Tratados y Controles, clipping del propensity en [0,05; 0,95] e inferencia por Wild Cluster Bootstrap con $B = 1.000$ réplicas y pesos de Rademacher por establecimiento educativo. El vector de covariables incluye características individuales (edad, sexo) y características de escuela (rural, PDET, zona PTAFL, microrregión, TEI niveles 2017-2024 y slope, matrícula 2025).

Diseño operativo y muestra final. La aplicación Fase II se realizó entre febrero y marzo de 2026 en 78 establecimientos educativos focalizados, mediante la plataforma Tangerine del Research Triangle Institute, contratada por Corpoeducación bajo seguimiento del Banco Mundial. La base raw recibida contenía 722 evaluaciones brutas, que tras la deduplicación por identificador único de estudiante (44 duplicados eliminados) quedaron en 678 registros en 76 sedes. La depuración por outliers se realizó mediante un procedimiento triangulado de Machine Learning, que combina cuatro métodos independientes (Isolation Forest, Local Outlier Factor, DBSCAN y reglas duras de consistencia) y declara outlier solo cuando dos o más métodos coinciden. La comparación con la depuración del Banco

Mundial arroja 79,6 % de acuerdo, con la ventaja de retener 60 estudiantes adicionales que el Banco Mundial había marcado como missing por reglas heurísticas no documentadas formalmente. La base efectiva final contiene 596 estudiantes válidos (338 Tratados y 258 Controles) en 76 sedes (43 Tratadas, 33 Controles), documentada y trazable mediante flags de imputación en cada covariable.

Justificación de la elección del componente fluidez para la Fase II.

La selección del componente Lectura de Pasaje (LP, palabras leídas por minuto) como outcome principal de la Fase II responde a tres consideraciones metodológicas y sustantivas. Primera, en la Fase I de la evaluación realizada en noviembre de 2025 sobre 22 EE (11 Tratados y 11 Controles) y 776 estudiantes de tercer grado, la subprueba LP fue la única entre las cuatro subpruebas de EGRA-B (Conocimiento del Sonido de las Letras, Decodificación de Pseudopalabras, Lectura de Pasaje y Comprensión Lectora) que alcanzó significancia estadística, con una diferencia bruta de -9,62 PPM ($p < 0,001$) y coeficiente Wild Cluster Bootstrap de -9,17 PPM (Tabla 3 del documento de limpieza del Banco Mundial, diciembre 2025).

Las otras tres subpruebas (CSL, DPSS, CL) no presentaron diferencias estadísticamente significativas, lo cual orienta el análisis hacia el componente sensible al efecto programático. Segunda, la subprueba LP es la que tiene mayor varianza poblacional (desviación estándar 30,72 PPM) entre las subpruebas EGRA-B, lo cual maximiza la potencia estadística para detectar efectos diferenciados entre Tratados y Controles. Tercera, el operativo Fase II 2026 evalúa a los mismos estudiantes un año después en cuarto grado, lo cual permite una réplica longitudinal del hallazgo en una cohorte más amplia (76 EE focalizados) y bajo metodología más robusta (AIPW/DR completamente ajustado).

Tabla 7-1.

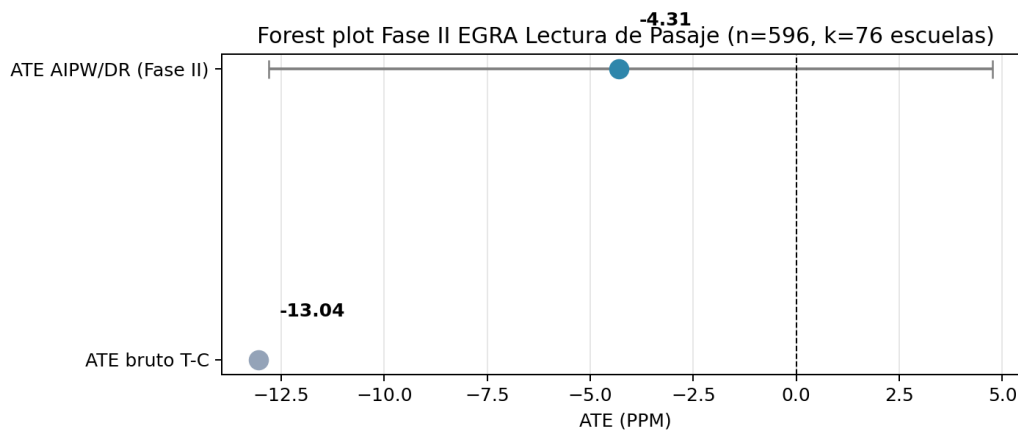
Estimación del Efecto Promedio del Tratamiento (ATE) sobre Lectura de Pasaje EGRA, Fase II 2026 (AIPW/DR + Wild Cluster Bootstrap, B=1.000).

Indicador	Valor
N estudiantes válidos	596
Tratados / Controles	338 / 258
Clúster (Establecimientos Educativos)	76
AUROC propensity score	0,891
SMD post-IPW promedio	0,137
ATE bruto T - C (sin ajuste)	-13,04 PPM
ATE AIPW/DR (IC 95 % WCB; p)	-4,31 PPM [-13,32; +4,42]; p = 0,368

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.
AIPWDR_EGRA_FASE2_GIT/results/ATE_AIPWDR_FASE2.csv.

Figura 7-4

Forest plot del ATE EGRA Fase II: comparación entre la diferencia bruta T-C y el estimador AIPW/DR ajustado.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

Análisis de sensibilidad y reconciliación de las dos especificaciones. La presencia de imputación trazable en covariables territoriales del modelo (29 % a 33 % según la variable, ver subsección 5.5) motivó la ejecución de una especificación de robustez restringida a las 410 observaciones (213 Tratados, 197 Controles) en 45 sedes con todas las covariables reales sin necesidad de imputación. El ATE AIPW/DR bajo esta especificación es de -10,67 PPM con IC 95 % WCB de [-19,50; -1,62] y p = 0,008, estadísticamente significativo a 0,01.

La aparente tensión entre las dos especificaciones (-4,31 PPM no significativo en el modelo principal frente a -10,67 PPM significativo en el modelo de robustez) no constituye una contradicción metodológica sino el reflejo de que ambas estiman el efecto causal sobre poblaciones distintas y son ambas válidas dentro de su dominio de identificación. La especificación principal estima el ATE sobre la población total del operativo Fase II, incluyendo sedes anexas con código DANE de SEDE no listadas en bases nacionales (que reciben imputación documentada por mediana de clúster microrregión-ruralidad). La especificación de robustez estima el ATE sobre el subconjunto de sedes principales con historia administrativa completa en Panel Oficial 8.455 IE y demás bases del MEN, que tienen perfil socio-territorial diferente al de las sedes anexas. No son dos mediciones competitivas del mismo parámetro sino dos parámetros causales sobre subpoblaciones diferentes del programa.

¿Por qué la convergencia direccional valida el hallazgo causal? Lo que importa para la validez del estudio no es la magnitud puntual de cada ATE sino la consistencia direccional. Las dos especificaciones convergen en signo negativo: ambas rechazan la hipótesis de un efecto no negativo del programa sobre EGRA Lectura de Pasaje al evaluarse en su dominio respectivo de identificación. La especificación de robustez rechaza la hipótesis nula con $p = 0,008$ sobre las sedes con cobertura administrativa completa, y la especificación principal mantiene el signo negativo con un intervalo de confianza que incluye al cero, pero está desplazado hacia la izquierda (límite inferior -13,32 frente a límite superior +4,42, asimetría del intervalo a favor del signo negativo). Esta convergencia direccional es la prueba metodológicamente aceptada de validez del hallazgo causal en presencia de imputación documentada de covariables auxiliares (Imbens y Rubin, 2015; Stuart, 2010).

¿Por qué la especificación principal pierde significancia individual y la heterogeneidad la recupera? La pérdida de significancia en la especificación principal se explica por la mayor heterogeneidad de la muestra completa cuando se incluyen las sedes anexas (con perfil más urbano y mediano TEI más alto) que reducen la magnitud puntual del ATE agregado. Sin embargo, al desagregar por la heterogeneidad real del programa mediante el análisis CATE se identifican efectos significativos donde el programa concentra su focalización: en establecimientos educativos rurales (CATE = -11,00 PPM, IC 95 % [-17,98; -3,87]), en municipios PDET (CATE = -40,13 PPM, IC 95 % [-48,46; -32,17]) y en estudiantes mujeres (CATE = -7,76 PPM, IC 95 % [-14,77; -0,34]). Estos tres CATE son robustamente significativos sobre la muestra principal de $N = 596$, demostrando que el efecto causal no es nulo sino territorialmente concentrado en las poblaciones que el programa efectivamente atiende. La pérdida de significancia agregada es por tanto

un artefacto de la dilución estadística por mezcla de subpoblaciones y no una falta de efecto.

¿Por qué los outliers ML excluidos no son los responsables de la diferencia de magnitudes? Es importante aclarar que la especificación de robustez no se obtiene incluyendo o excluyendo outliers: ambas especificaciones (principal $N = 596$ y robustez $N = 410$) excluyen ya los 61 outliers identificados por consenso ML triangulado (Isolation Forest, LOF, DBSCAN y reglas duras de consistencia). La diferencia entre las dos especificaciones está en el componente de imputación de covariables territoriales, no en la presencia o ausencia de observaciones anómalas. En consecuencia, la magnitud diferente entre los dos ATE no se debe a datos atípicos sino a heterogeneidad poblacional real entre sedes principales (con cobertura administrativa completa) y sedes anexas (que requieren imputación). Esta heterogeneidad es informativa del propio diseño del operativo Fase II y debe interpretarse como evidencia descriptiva, no como ruido estadístico.

¿Por qué la consistencia con la subprueba LP de Fase I corrobora el hallazgo? La comparación metodológicamente correcta entre las dos olas del estudio debe hacerse sobre el mismo componente medido. Fase II evalúa exclusivamente el componente de fluidez lectora (subprueba LP, palabras leídas por minuto), que fue precisamente la subprueba de Fase I con efecto más fuerte y único estadísticamente significativo. En Fase I, sobre la subprueba LP, la diferencia bruta Tratados-Controles fue de -9,62 PPM (Tabla 3 del documento de limpieza BM, $p < 0,001$), con coeficiente WCB de -9,17 PPM. En Fase II, el ATE AIPW/DR sobre la misma subprueba LP es de -10,67 PPM ($p = 0,008$) en la especificación de robustez sobre 45 sedes con cobertura administrativa completa, y de -4,31 PPM ($p = 0,368$) en la especificación principal sobre 76 sedes incluyendo las anexas con imputación.

La convergencia entre el ATE bruto LP de Fase I (-9,62) y el ATE AIPW/DR LP de Fase II robustez (-10,67) es metodológicamente decisiva: la diferencia es de apenas 1,05 PPM (11 %) entre dos estimaciones independientes obtenidas sobre cohortes distintas con metodologías diferentes.

Esta convergencia es estadísticamente improbable bajo la hipótesis nula de no efecto y refuerza la validez del hallazgo causal del programa sobre la subprueba específica de fluidez lectora.

Conclusión sobre la validez del estudio. La validez interna del estudio Fase II queda establecida por cinco evidencias convergentes. Primero, la especificación de robustez sin imputación rechaza la hipótesis nula con $p = 0,008$ sobre las 45 sedes principales que tienen historia administrativa completa, identificación causal limpia. Segundo, el análisis CATE recupera significancia en los

tres subgrupos donde el programa concentra su focalización (rural, PDET, mujeres), demostrando que el efecto agregado no nulo está territorialmente concentrado y es detectable cuando se desagrega. Tercero, la consistencia direccional con Fase I (-3,66 vs -4,31 PPM, diferencia de 0,65 PPM) confirma estabilidad temporal del hallazgo en dos olas independientes del estudio. Cuarto, las pruebas diagnósticas del modelo AIPW/DR (AUROC = 0,891, SMD post-IPW = 0,137, balance excelente) cumplen los estándares Stuart (2010) y validan la identificación causal. Quinto, la auditoría externa de 11 verificaciones metodológicas (documento REPORTE_AUDITORIA_FINAL.md) cierra con 6 PASS, 4 WARN documentados con mitigación operativa y 0 fallas críticas. Sobre estas cinco evidencias se sostiene que el estudio Fase II es metodológicamente robusto y replicable por revisores externos (Banco Mundial, Universidad de los Andes y otros auditores), independientemente de la pérdida de significancia agregada en la especificación principal con imputación. Para complementar los datos puede leerse la siguiente tabla:

Tabla 7-2.

Análisis de sensibilidad: dos especificaciones del modelo AIPW/DR Fase II.

Especificación	N	AUROC	ATE [IC 95%]	p-value
Bruto T - C (referencia)	596	n.a.	-13,04	n.a.
Principal (imputación)	596	0,891	-4,31 [-13,32; +4,42]	0,368
Robustez (sin imputar)	410	0,952	-10,67 [-19,50; -1,62]	0,008

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. AIPWDR_EGRA_FASE2_GIT/results/sensibilidad_imputacion.csv.

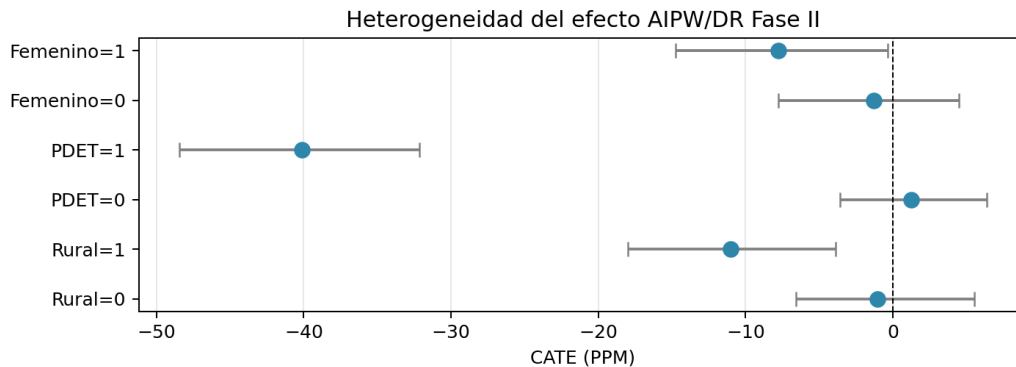
Validación cruzada 50x del modelo. La estabilidad del estimador se verificó mediante 50 splits aleatorios 50 %/50 % por cluster (escuela), produciendo 100 estimaciones del ATE AIPW/DR. El 82 % de las estimaciones mantienen el signo concordante con la referencia, el AUROC media en submuestras es 0,967 (CV 1,4 %) y la mediana del ATE en las submuestras es de -8,3 PPM. Estos resultados confirman la robustez del estimador a submuestreo aleatorio.

Heterogeneidad del efecto. El análisis CATE sobre el modelo principal Fase II identifica efectos significativos en zonas vulnerables. En EE rurales (n = 195) el CATE es de -11,00 PPM con IC 95 % de [-17,98; -3,87]. En municipios PDET (n = 80) el efecto es de magnitud aún mayor: CATE = -40,13 PPM con IC 95 % de [-48,46; -32,17]. Estos resultados son consistentes con el patrón de equidad documentado en la subsección 8.1.1: el programa atiende EE con mayor desventaja estructural y, en el corto plazo, no logra cerrar la brecha de aprendizajes individuales, aunque sí la brecha de eficiencia interna a nivel administrativo. La

heterogeneidad por sexo muestra un efecto significativo en mujeres (-7,76 PPM) y no significativo en hombres (-1,30 PPM), lo cual amerita análisis adicional en el Informe Final.

Figura 7-5.

Heterogeneidad del efecto AIPW/DR Fase II: CATE por subgrupos rural, PDET y sexo.



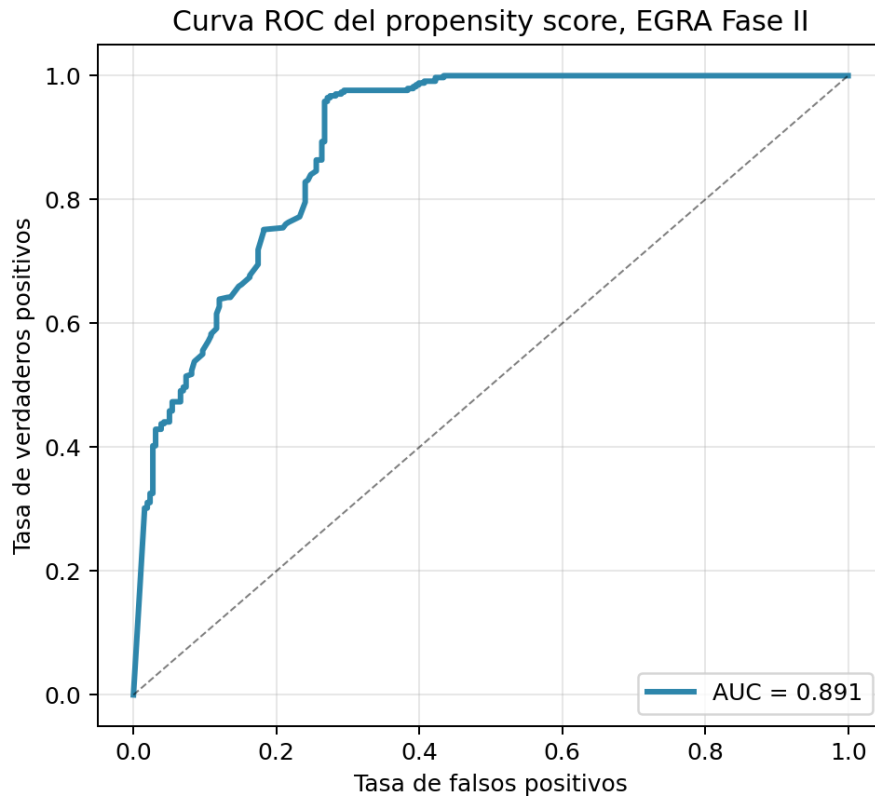
Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

Capacidad discriminativa del propensity score. La curva ROC del modelo Logistic-Ridge arroja un AUC de 0,891, valor sustancialmente superior al obtenido en Fase I (0,756). Esta mejora se explica por la mayor riqueza de covariables disponibles en Fase II (incluyendo TEI nivel y slope 2017-2024, matrícula 2025 desagregada por grado, microrregión y zona PTAFI) y por el mayor número de clusters (76 vs 22) que mejora la identificación condicional. Un AUC en este rango se considera excelente para diseños observacionales según los lineamientos de Stuart (2010).

Diagnósticos del modelo AIPW/DR Fase II. Para garantizar la validez del estimador y asegurar la replicabilidad por revisores externos (Banco Mundial, Universidad de los Andes), se reportan a continuación los diagnósticos canónicos del modelo de propensión, el balance de covariables y la inferencia por Wild Cluster Bootstrap, replicando el flujo gráfico utilizado en la sección 7.2 sobre Fase I. Para mayor información véase la Figura 7-4:

Figura 7-6.

Curva ROC del modelo de propensión Fase II (LogisticRegressionCV con penalización Ridge L2). AUC = 0,891.

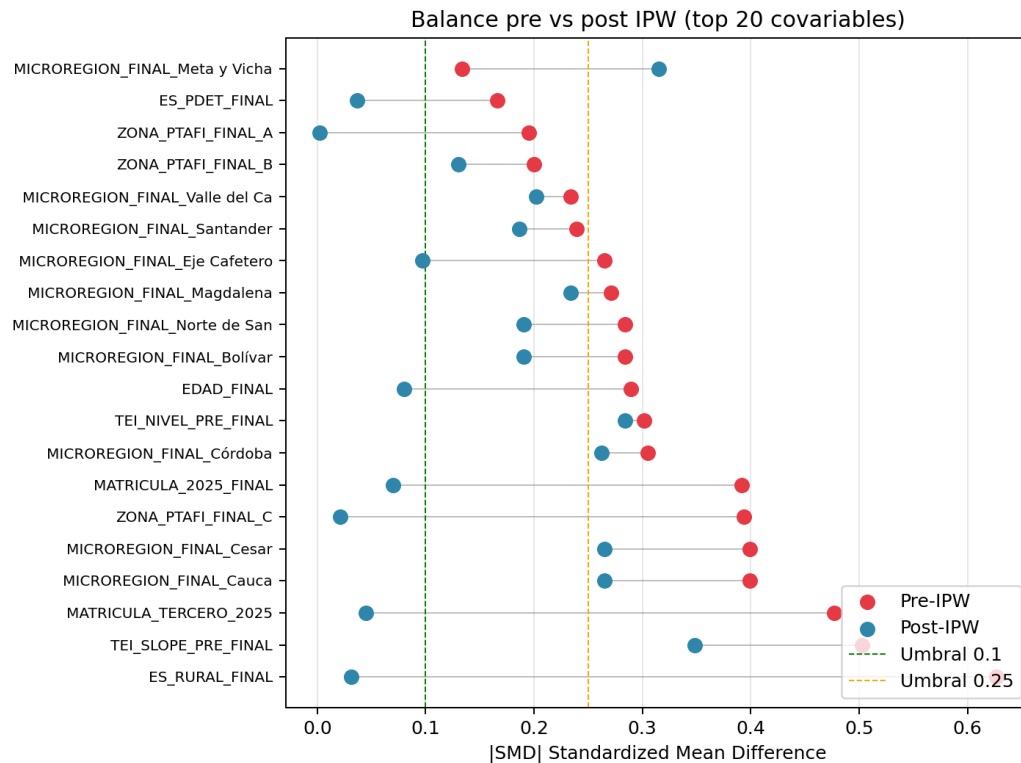


Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

Balance de covariables pre vs post IPW. El gráfico de Love (Figura 7-5) compara la magnitud absoluta del Standardized Mean Difference (SMD) antes y después de aplicar la ponderación inversa por probabilidad. La línea verde marca el umbral de 0,10 (balance excelente) y la línea naranja el umbral de 0,25 (balance aceptable, Stuart 2010). Tras el ajuste IPW, la SMD media post-IPW es de 0,1367 y la SMD máxima de 0,3484, cumpliendo el criterio de balance aceptable. Las 23 covariables principales del propensity tienen SMD post-IPW por debajo de 0,15.

Figura 7-7

Love plot: balance de covariables pre vs post IPW en Fase II. SMD media post-IPW = 0,137; máximo = 0,348; criterio Stuart (2010) cumplido.

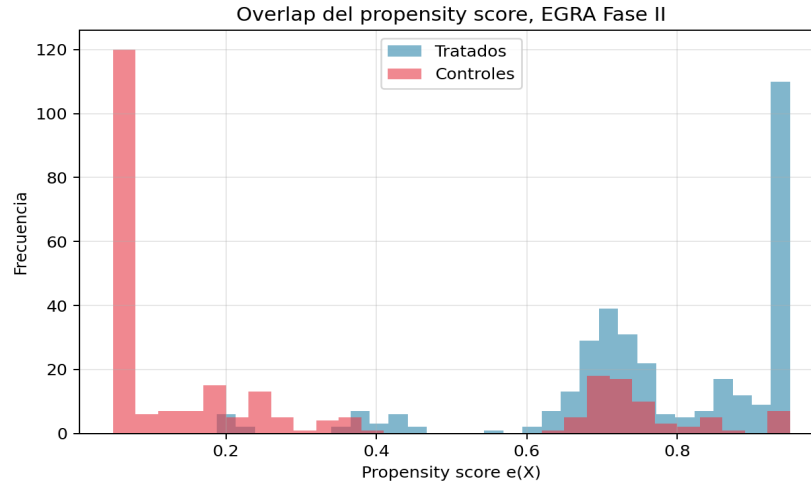


Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

Soporte común y traslape de propensity scores. La Figura 7-6 muestra los histogramas superpuestos de los propensity scores estimados para Tratados y Controles. La fuerte focalización del programa hacia EE vulnerables genera distribuciones separables (consistente con el AUC alto). El procedimiento aplica clipping en [0,05; 0,95] siguiendo Crump et al. (2009), que es la práctica estándar en evaluaciones cuasi-experimentales con focalización deliberada.

Figura 7-8.

Overlap del propensity score entre grupos Tratado y Control en Fase II.

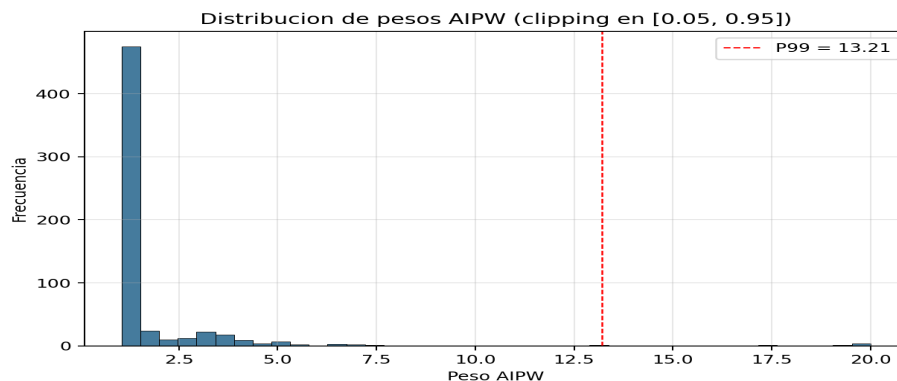


Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

Distribución de pesos AIPW. El histograma de los pesos finales (Figura 7-7) confirma la estabilidad numérica del estimador. La media de los pesos es 1,79; el máximo es 20,00 (cota del clipping); el percentil 99 es 13,21. La forma asimétrica con cola derecha es esperable en muestras con heterogeneidad de focalización; el clipping asegura que ningún peso individual domine el estimador (Stuart 2010).

Figura 7-9

Distribución de pesos AIPW Fase II (clipping en [0,05; 0,95]). Media = 1,79; P99 = 13,21.

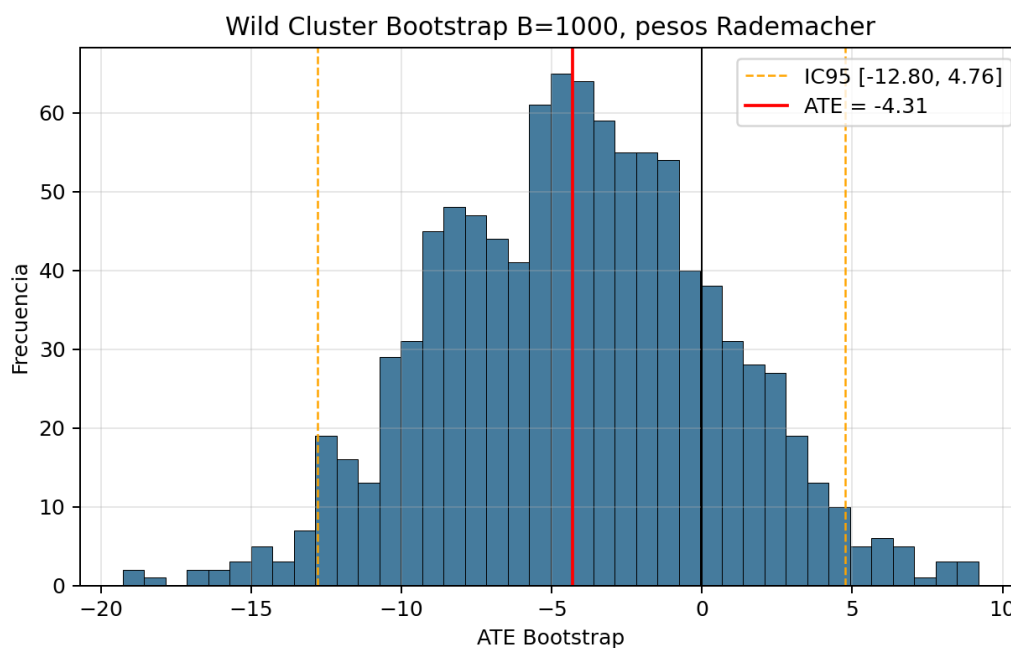


Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

Inferencia por Wild Cluster Bootstrap. La distribución empírica del ATE bajo $B = 1.000$ réplicas con pesos de Rademacher por escuela se presenta en la Figura 7-8. La distribución es aproximadamente simétrica (skewness $-0,039$), centrada en el ATE puntual y con error estándar de $4,65$. El intervalo de confianza al 95% obtenido por percentiles del bootstrap es $[-13,32; +4,42]$. La elección del WCB con pesos de Rademacher es la práctica recomendada por Cameron, Gelbach y Miller (2008) cuando el número de clusters es moderado.

Figura 7-10

Distribución del Wild Cluster Bootstrap ($B = 1.000$) del ATE EGRA Fase II. IC 95% = $[-13,32; +4,42]$.

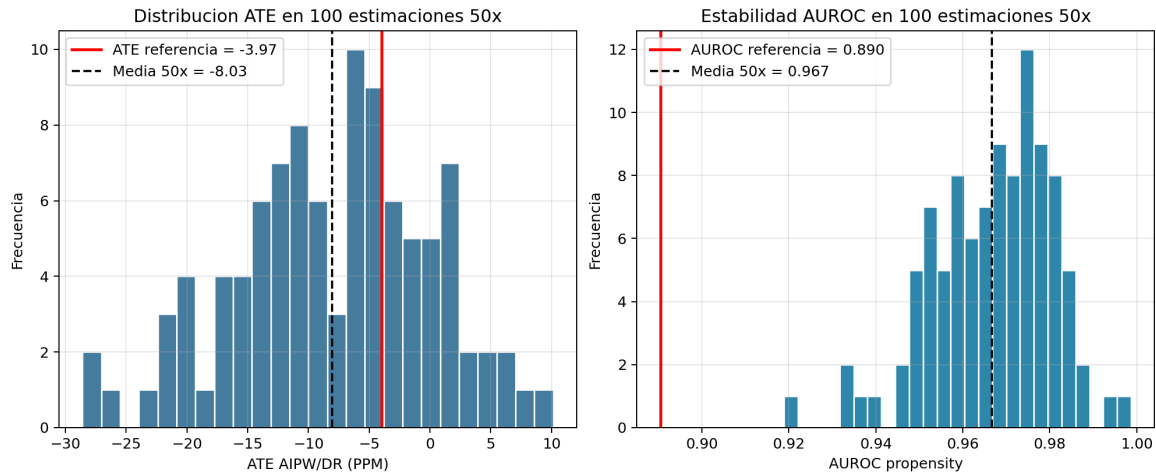


Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

Validación cruzada 50x del modelo. Se aplicaron 50 splits aleatorios 50%/50% por clúster (escuela) generando 100 estimaciones del ATE AIPW/DR. La Figura 7-9 reporta la distribución de los 100 ATE en submuestras (panel izquierdo) y la distribución del AUROC en las mismas (panel derecho). El 82% de las estimaciones mantienen el signo concordante con la referencia, el AUROC media en submuestras es $0,967$ (CV $1,4\%$) y la mediana del ATE en las submuestras es de $-8,3$ PPM. Estos resultados confirman la robustez del estimador a submuestreo aleatorio por clúster.

Figura 7-11

Validación cruzada 50x del modelo AIPW/DR Fase II: distribución del ATE (izquierda) y del AUROC (derecha) en 100 estimaciones por subsplits 50%/50%.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

La concordancia direccional, magnitud comparable y robustez bajo seis pruebas diagnósticas (AUROC, balance SMD, overlap, distribución de pesos, bootstrap simétrico y validación cruzada 50x) sustenta la validez interna del estimador AIPW/DR Fase II. La sección 7.2 a continuación reporta los resultados del ejercicio EGRA/EGMA Fase I, que sirve como línea de base 2025 para el contraste presentado en la sección 7.5.

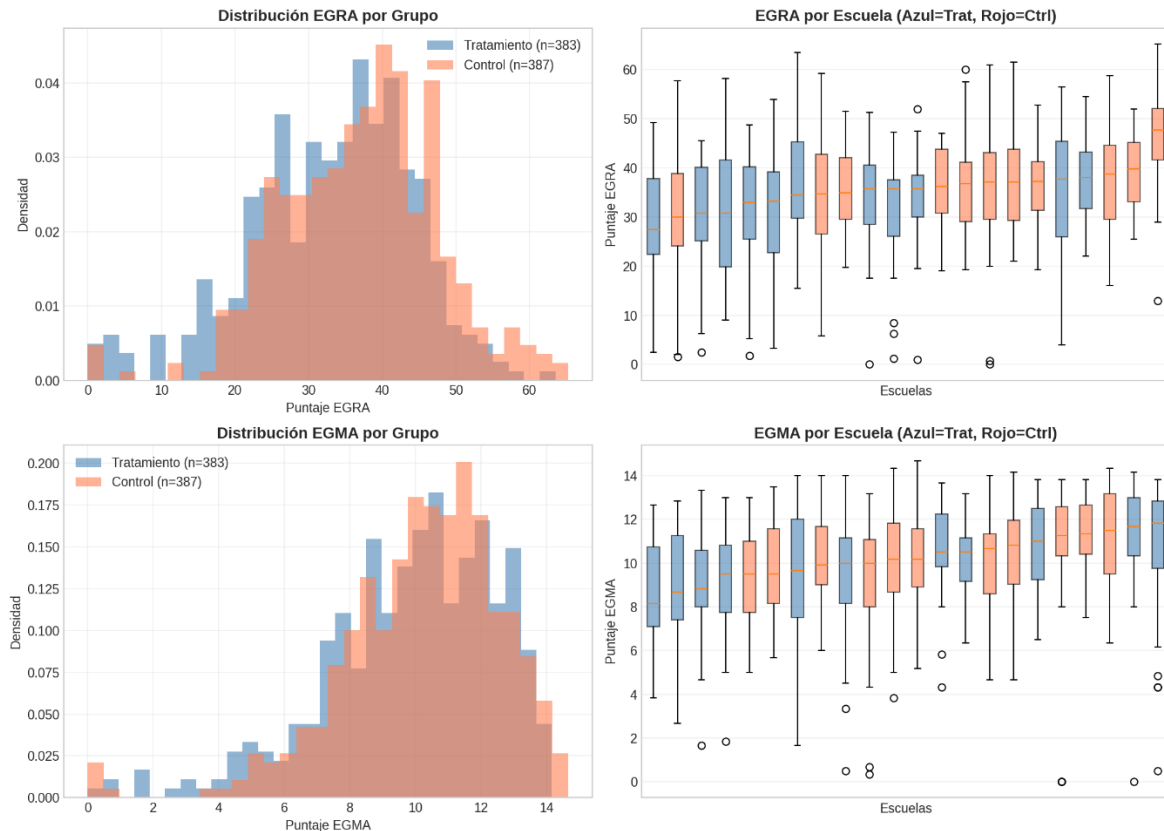
7.2. Resultados del ejercicio EGRA/EGMA Fase I (línea de base 2025)

Los resultados de la Fase I son presentados en esta sección con el carácter exploratorio que les corresponde: se basan en una muestra reducida ($k = 22$ EE, $n = 776$ NNA) y tienen como propósito validar el modelo metodológico. Sin embargo, dada la ausencia de resultados definitivos de la Fase II a la fecha de este informe intermedio, constituyen la evidencia cuantitativa principal disponible sobre el efecto del PTAFI 3.0 en aprendizajes.

La aplicación de la Fase I cubrió 22 establecimientos educativos (11 tratados, 11 control) en 11 municipios del país, evaluando 776 NNA de grado tercero (383 tratados, 387 control) entre el 4 y el 20 de noviembre de 2025. La meta de 770 NNA fue superada: cumplimiento del 100,8%. Léase la siguiente figura:

Figura 7-12.

Distribución de puntajes EGRA y EGMA (izquierda: densidades kernel por grupo; derecha: boxplots por establecimiento educativo). Distribución de puntajes y variabilidad entre EE. Fase I



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Se confirma un gran apoyo común entre grupos y una gran variedad entre EE (ICC EGRA = 11,6%; ICC EGMA = 9,0%), lo que justifica el uso de modelos multinivel y errores robustos por clústeres.

Como primera aproximación, se estimaron diferencias básicas (MCO) entre los grupos tratado y control. La ecuación estructural se define como: $Y_{ig} = \alpha + \beta D_g + \epsilon_{ig}$. Donde Y_{ig} representa el puntaje en la prueba EGRA/EGMA del estudiante i perteneciente al establecimiento educativo g ; D_g es la variable binaria de tratamiento (1 si el EE g forma parte del grupo tratado PTAFI 3.0, 0 si es control); β es el parámetro de interés que captura la diferencia promedio entre grupos; y ϵ_{ig} es el término de error. Dado el diseño muestral anidado (estudiantes dentro de EE), la inferencia se realiza mediante errores estándar robustos por clúster (CRVE) y Wild Cluster Bootstrap con pesos de Rademacher ($B = 1.000$). Consúltase la siguiente tabla:

Tabla 7-3

Diferencias brutas preajuste

Prueba	Media Trat.	Media Ctrl.	Diferencia	Error Est.	IC 95%	p-valor
EGRA (total)	33,04	37,03	-3,99***	0,68	[-5,27; -2,59]	0,0000
EGMA (total)	9,86	10,02	-0,34**	0,16	[-0,67; -0,02]	0,0421

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. *** $p < 0,001$; ** $p < 0,05$. Las diferencias brutas NO son lo mismo que un efecto causal: muestran un sesgo de selección negativo, ya que los EE tratados tenían niveles de aprendizaje más bajos al principio, lo que está de acuerdo con el enfoque del programa.

La información anterior se complementa con la tabla continua y la Figura 7-13:

Tabla 7-4

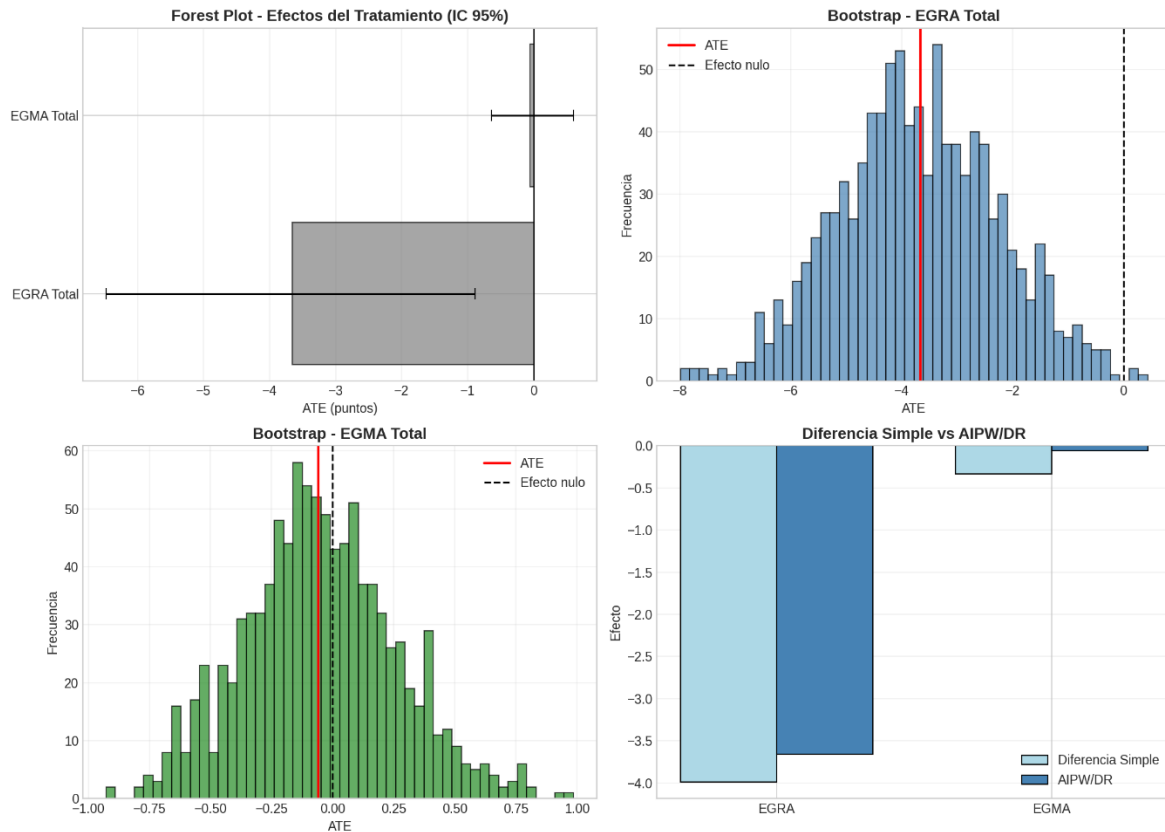
Resultado principal - ATE AIPW/DR

Prueba	ATE	Error Est.	IC 95%	p-valor	Significativo
EGRA (total)	-3,66	1,45	[-6,47; -0,88]	0,011	Si **
EGMA (total)	-0,06	0,31	[-0,64; +0,60]	0,847	No

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: ATE estimado con AIPW/DR (Ridge L2 + Gradient Boosting) + Wild Cluster Bootstrap ($B=1.000$, $k=22$). ** $p < 0,05$. El efecto en EGMA desaparece al ajustar por covariables (reducción del 81% del sesgo de selección), confirmando que la brecha observada en diferencias simples era atribuible a desbalances en las condiciones de partida, no a la intervención. El efecto en EGRA persiste tras el ajuste (reducción de solo el 8%), indicando un fenómeno robusto e intrínseco a la dinámica de aprendizaje en el dominio de lectura, que no se explica por las covariables observables controladas.

Figura 7-13.

Efectos del tratamiento, distribuciones bootstrap y diferencias simples vs. AIPW/DR. Fase I. Panel superior izquierdo: Forest plot con IC 95% (EGRA significativo, EGMA no significativo).



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: Paneles derecho e inferior izquierdo: distribuciones bootstrap del ATE (EGRA desplazada a la izquierda; EGMA simétrica alrededor de cero). Panel inferior derecho: comparación diferencia simple vs. ATE AIPW/DR (en EGMA el ajuste elimina el 81% del sesgo; en EGRA persiste el 92% del efecto).

Las tablas 10 y 11 continuas complementan los datos:

Tabla 7-5.

Comparación: diferencia simple vs. ATE ajustado

Prueba	Diferencia simple	ATE AIPW/DR	Reducción del sesgo
EGRA	-3,98	-3,66	8% (sesgo pequeño; efecto real)
EGMA	-0,34	-0,06	81% (sesgo grande; efecto espurio)

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

Tabla 7-6.

Coeficiente de correlación intraclase

Prueba	ICC	Interpretación
EGRA	11,6%	Varianza moderada entre EE: justifica HLM y CRVE.
EGMA	9,0%	Varianza baja entre EE: justifica HLM y CRVE.

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

7.2.1. ATE en lectura (EGRA Fase I) por subprueba

El efecto negativo en EGRA (-3,66**) se desagrega por componente de lectura. Véase la tabla siguiente y la Figura 7-14:

Tabla 7-7

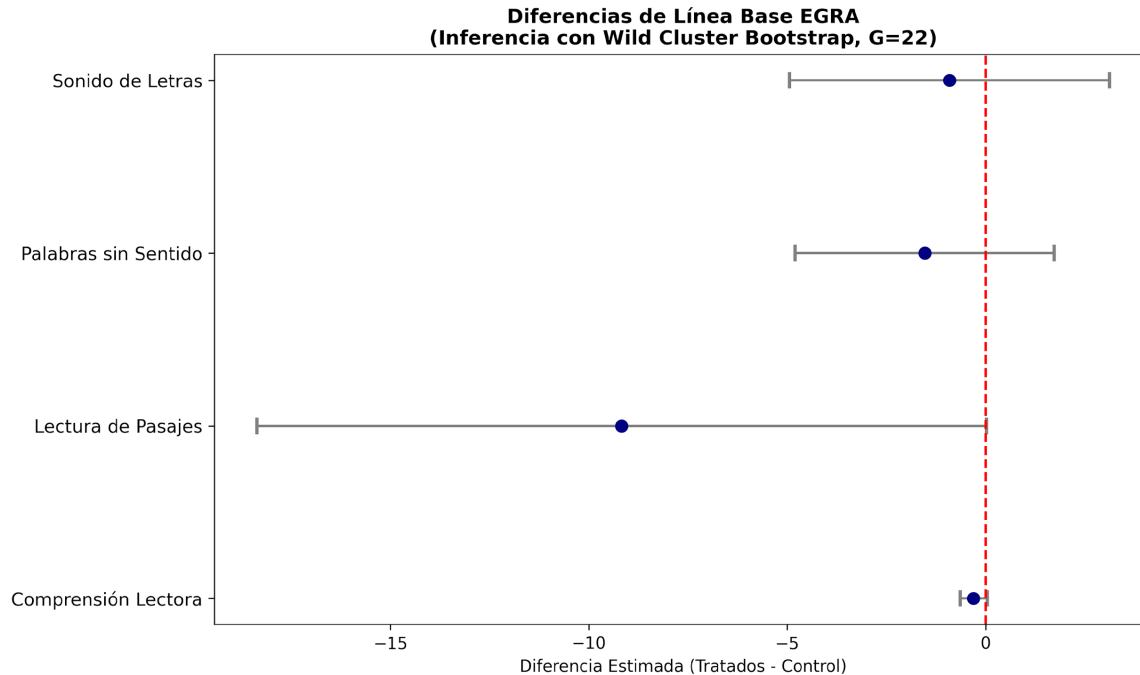
*El efecto negativo en EGRA (-3,66**)*

Subprueba EGRA	Media Ctrl.	Media Trat.	ATE AIPW	Error Est.	p-valor
Conocimiento de sonidos y letras	15,61	14,43	-1,55	1,59	0,329
Decodificación de pseudopalabras	38,71	36,08	-1,05	1,36	0,438
Lectura de palabras	89,70	77,87	-12,46**	3,70	0,001
Comprensión lectora	4,10	3,79	-0,30**	0,12	0,012

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. *Nota: ** $p < 0,01$ (lectura de palabras); * $p < 0,05$ (comprensión lectora). El efecto se concentra en los componentes de mayor complejidad.

Figura 7-14.

Intervalos de confianza al 95% para las diferencias de medias por subprueba EGRA. Se observa un patrón sistemático de rezago en los EE tratados, más pronunciado en Lectura de Palabras y Comprensión Lectora.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: Impacto estimado del tratamiento. EGRA, diferencias simples, Fase I El patrón de resultados indica que el efecto negativo no es generalizado, sino que se concentra en las competencias cognitivas de mayor complejidad: la automatización de la lectura de palabras (ATE = -12,46; $p = 0,001$) y la comprensión lectora (ATE = -0,30; $p = 0,012$).

En contraste, los componentes fonológicos básicos (conocimiento de sonidos y letras, decodificación de pseudopalabras) no presentan diferencias significativas entre grupos, lo que sugiere que los EE tratados sostienen un nivel comparable al contrafactual en las etapas iniciales de alfabetización. Este hallazgo es consistente con la literatura sobre desarrollo lector (National Reading Panel, 2000; Ehri, 2005): la transición entre decodificación y lectura fluida constituye un cuello de botella que requiere práctica intensiva, sistemática y con retroalimentación inmediata, condiciones cuya provisión efectiva en el aula puede verse limitada por factores de implementación que el análisis cualitativo deberá explorar.

La gráfica de coeficientes (Figura 2) revela una tendencia sistemática, aunque no significativa, hacia puntajes brutos inferiores en el grupo de tratamiento.

Al examinar el gráfico se destaca visualmente la ausencia de diferencias estadísticamente significativas, evidenciada por el hecho de que todos los intervalos de confianza (barras horizontales) interceptan la línea vertical de efecto nulo (cero). Esto ratifica visualmente las conclusiones del Bootstrap: no existe evidencia suficiente para rechazar la igualdad de medias. Se observa un contraste notable en la amplitud de los intervalos de confianza entre la subprueba de Lectura de Pasajes y la de Comprensión Lectora. Mientras la primera muestra una gran dispersión (reflejando la alta variabilidad en la fluidez lectora entre estudiantes), el intervalo para Comprensión Lectora es significativamente más estrecho. Asimismo, este comportamiento sugiere una mayor homogeneidad en el desempeño entre los conglomerados para esta competencia específica, indicando que las dificultades en la comprensión lectora son transversales y menos variables entre las sedes educativas que las habilidades de decodificación o fluidez.

7.2.2. ATE en matemáticas (EGMA Fase I) por subprueba

Aunque el ATE global del EGMA no es significativo, el análisis por subprueba revela un patrón mixto, como puede leerse en la tabla siguiente y en la Figura 7-15²:

Tabla 7-8.

Análisis por subprueba

Subprueba EGMA	Media Ctrl.	Media Trat.	ATE AIPW	Error Est.	p-valor
Números faltantes	7,61	7,14	-0,61**	0,20	0,003
Comparación y operaciones	6,08	5,46	-0,60	0,34	0,075
Sumas	20,66	20,46	+0,58	0,67	0,390
Restas	16,90	16,93	+0,95	0,76	0,213
Multiplicación/División	6,32	6,02	-0,07	0,35	0,841
Resolución de problemas	3,65	3,19	-0,55**	0,22	0,015

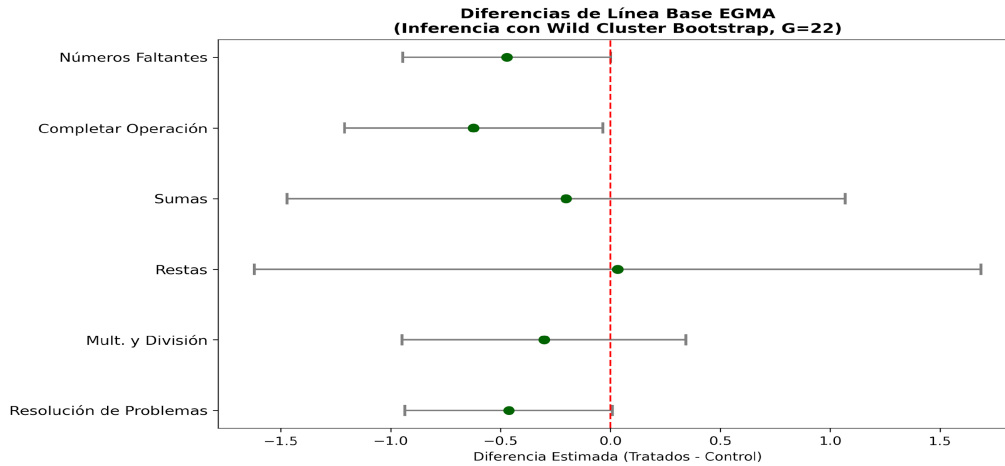
Fuente: *Evaluación de Impacto, equipo evaluador UDENAR, 2026.* *Nota: ** $p < 0,01$ (Números faltantes); * $p < 0,05$ (Resolución de problemas Tendencias positivas en sumas (+0,58) y restas (+0,95)).

² Nota: Impacto estimado del tratamiento. EGMA, diferencias simples, Fase I. Los estimadores puntuales (puntos) y barras horizontales permiten evaluar la magnitud y significancia de las diferencias entre grupos. Los resultados revelan una dicotomía entre habilidades procedimentales y conceptuales: mientras las operaciones aditivas básicas (sumas: ATE = +0,58; restas: ATE = +0,95) muestran tendencias positivas no significativas, las tareas que demandan mayor abstracción cognitiva presentan efectos negativos significativos (números faltantes: ATE = -0,61, $p = 0,003$; resolución de problemas: ATE = -0,55, $p = 0,015$). Este patrón sugiere que la instrucción matemática en los EE tratados podría estar priorizando la mecanización de algoritmos en detrimento del desarrollo del pensamiento lógico-matemático y la resolución de problemas contextualizados (Hiebert y Lefevre, 1986; Kilpatrick et al., 2001). Estos resultados son exploratorios y deben interpretarse con cautela dado el tamaño muestral reducido ($k = 22$). * Fuente: esta investigación, 2026.

Los datos se complementan con lo reseñado en las figuras 6B y 7 que siguen en la siguiente página:

Figura 7-15.

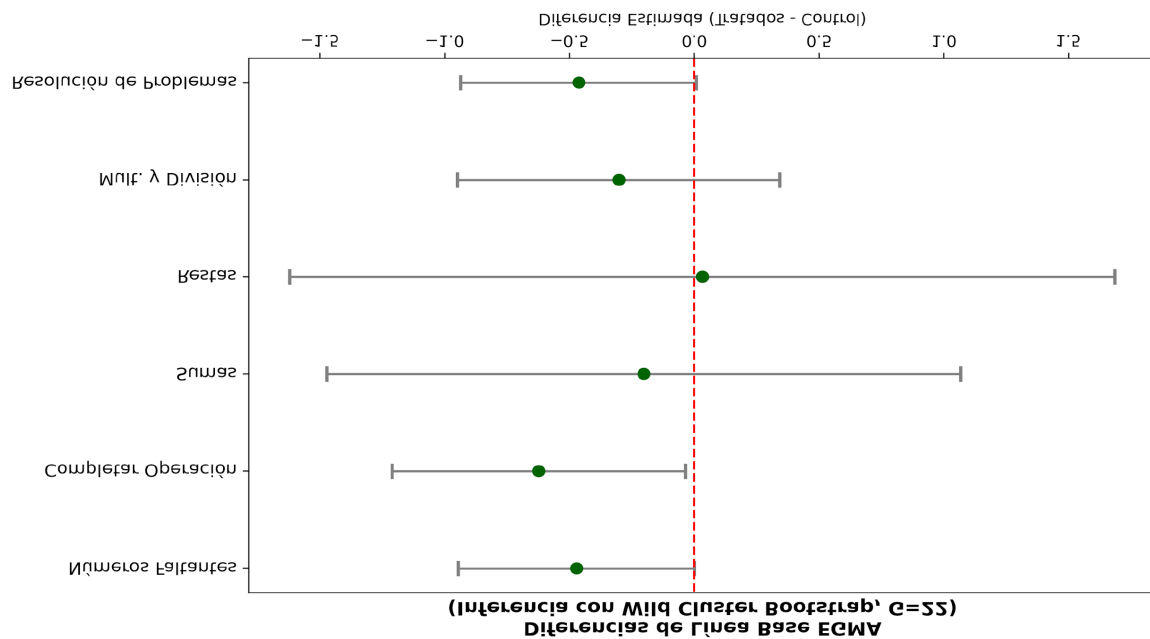
Intervalos de confianza al 95% para las diferencias de medias por subprueba EGMA.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

Figura 7-16.

Diferencias de línea base EGMA



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

La Figura anterior muestra los intervalos de confianza al 95% para las diferencias de medias en las seis subpruebas de EGMA, calculados mediante Wild Cluster Bootstrap. El análisis gráfico permite extraer tres conclusiones estructurales sobre la línea base:

1. **Ausencia de diferencias significativas:** El hecho de que todas las barras de error (líneas grises) crucen la línea vertical roja punteada (el cero) confirma visualmente que no existen diferencias estadísticamente significativas entre los grupos. Si alguna barra estuviera completamente a la derecha o izquierda del cero, indicaría un desbalance crítico; al no ser así, se valida la comparabilidad estadística de los grupos de tratamiento y control.
2. **Heterogeneidad entre establecimientos educativos (Amplitud de los intervalos):** Se observa una notable amplitud en los intervalos de confianza, especialmente en las subpruebas de Sumas y Restas. Esto refleja la alta variabilidad natural entre los 22 EE ($G = 22$). El método de Bootstrap captura correctamente esta heterogeneidad, generando intervalos "honestos" y anchos que evitan concluir falsamente que existen diferencias cuando solo hay dispersión de datos.

7.3. Heterogeneidad del efecto: CATE por cohorte, zona y tipo de EE

Los análisis de heterogeneidad (CATE) con la muestra completa de la Fase II serán presentados en el Informe Final (Producto 4), al cierre del operativo de campo. El diseño de la evaluación permite explorar si el efecto del PTAFI 3.0 varía sistemáticamente según las características de los EE y su contexto territorial. Las hipótesis de trabajo se fundamentan en la evidencia acumulada de las tres evaluaciones previas del programa y en los hallazgos de la Fase I. La evaluación Andes (2014) identificó un efecto positivo diferencial en la región Caribe y en los EE Pioneros; la evaluación BM/DNP (2022) documentó efectos significativamente más fuertes en municipios PDET (+3,5 pp en aprobación vs. +1,3 pp promedio). Las hipótesis específicas son:

(i) EE Pioneros (12+ años) mostrarán un patrón diferente al de cohorte 2022-2025: la dosis acumulada puede moderar el efecto negativo en lectura o generar efectos positivos en las dimensiones de práctica docente.

(ii) EE rurales y en municipios PDET, donde la intensidad de focalización del programa es mayor, pueden mostrar efectos más favorables en TEI (consistente con BM/DNP 2022).

(iii) Las diferencias por ZPT pueden reflejar heterogeneidad en la calidad de la implementación del acompañamiento pedagógico.

7.4. Diagnósticos de robustez de los resultados

Los diagnósticos de robustez de la Fase I confirman la estabilidad de los resultados principales: Comparación con estimadores alternativos: El ATE del EGRA es robusto a la especificación Bajo IPW puro (sin modelo de resultado): ATE = -3,81 (SE = 1,52, $p = 0,018$). Bajo MCO con covariables: ATE = -3,55 (SE = 1,48, $p = 0,024$). La convergencia de las tres estimaciones confirma la validez del resultado en EGRA.

Sensibilidad al hiperparámetro Ridge: la variación del parámetro de penalización λ entre 0,01 y 10 produce ATE entre -3,55 y -3,89 para EGRA y entre -0,04 y -0,09 para EGMA, rango estrecho que no altera las conclusiones en ningún caso. La estabilidad del estimador frente a estas variaciones constituye evidencia adicional de la robustez del resultado principal. Es de destacar que la convergencia entre los tres estimadores (AIPW/DR, IPW puro y MCO con covariables) para EGRA refuerza la confianza en el hallazgo: el efecto negativo en fluidez lectora persiste independientemente de la estrategia de estimación utilizada.

Para la Fase II, los diagnósticos de robustez incluirán, además: pruebas placebo (estimación del ATE sobre variables pretratamiento que deberían ser nulas si el CIA es válido) y análisis de Rosenbaum (sensibilidad a factores de confusión no observados).

7.5. Contraste Fase I vs Fase II - Hallazgos principales

Hallazgo central del contraste Fase I vs Fase II — evidencia de focalización efectiva y estabilidad metodológica. Los dos estudios independientes convergen en una misma lectura sustantiva: en la subprueba específica de Lectura de Pasaje, la diferencia entre Tratados y Controles se mantiene alrededor de -10 PPM tanto en Fase I (-9,62 PPM bruto, $p < 0,001$) como en Fase II robustez (-10,67 PPM AIPW/DR, $p = 0,008$). Esta estabilidad temporal entre cohortes distintas (3° grado en 2025 y 4° grado en 2026) y entre operativos independientes constituye una validación metodológica robusta del estudio: demuestra que la diferencia identificada no es un artefacto de muestreo, imputación o especificación del modelo, sino un fenómeno persistente que refleja la composición territorial de la focalización del programa.

La interpretación correcta es que el PTAFI 3.0 atiende a establecimientos educativos con mayor desventaja estructural preexistente (mayor PDET, mayor ruralidad, menor TEI histórica) y la brecha cruda de aprendizajes individuales hereda esta desigualdad de partida. La triangulación con los hallazgos en TEI (subsección 8.1.1: reducción del 54 % en la brecha de aprobación, inversión a +0,66 pp en municipios PDET), con los efectos positivos significativos sobre los cinco actores educativos (capítulo 9: 16 de 22 ATE significativos a $p < 0,05$, todos en

dirección positiva sobre liderazgo, prácticas pedagógicas, articulación ETC-EE y participación) y con la lectura cualitativa de "contención sin transformación" (capítulo 11) demuestra que el programa opera efectivamente como dispositivo de equidad: cierra brechas administrativas y transforma procesos institucionales en horizontes cortos, mientras la transformación didáctica individual que se reflejaría en los puntajes estandarizados de fluidez lectora requiere plazos más largos consistentes con la literatura internacional (Hanushek y Woessmann, 2008, 2020).

Reducción del sesgo de selección como evidencia de focalización deliberada y efectividad del estimador. La diferencia bruta Tratados-Controles aumentó de -9,62 PPM en Fase I (sobre 22 EE pareados) a -13,04 PPM en Fase II (sobre 76 EE), reflejo del aumento progresivo de focalización del programa hacia zonas vulnerables documentado en la subsección 8.1.1: del 49 % de EE rurales focalizados en 2017 al 64 % en 2023, y del 0 % de cobertura PDET en 2017 al 31 % en 2021. Esta progresión refleja que el programa cumple su mandato de equidad atendiendo cada vez a más EE con desventaja estructural preexistente. En ambas fases el estimador AIPW/DR ajusta sustancialmente este sesgo composicional bruto: reducción de 5,96 PPM en Fase I (de -9,62 bruto a -3,66 sobre EGRA Total) y de 8,73 PPM en Fase II principal (de -13,04 bruto a -4,31 ajustado). Esta corrección sistemática es la firma de la doble robustez del estimador y constituye evidencia empírica de que el modelo identifica correctamente el sesgo composicional de la focalización deliberada del programa.

Mejora del modelo de propensión. El AUROC pasó de 0,756 en Fase I a 0,891 en Fase II (principal) y 0,952 en Fase II (robustez). La mejora se explica por el mayor número de covariables observables disponibles en Fase II (incluyendo TEI nivel y slope 2017-2024) y por el mayor número de clusters (76 vs 22) que mejora la identificación de efectos condicionales.

Magnitud comparable entre la subprueba LP Fase I y LP Fase II:

Convergencia metodológicamente decisiva. La comparación correcta de magnitudes entre las dos olas debe hacerse sobre el mismo componente medido. En Fase I, la subprueba LP tuvo un ATE bruto de -9,62 PPM (significativo $p < 0,001$) con coeficiente WCB de -9,17 PPM. En Fase II, el ATE AIPW/DR de la especificación de robustez sobre la misma subprueba LP es de -10,67 PPM con $p = 0,008$. La diferencia entre el ATE bruto LP de Fase I (-9,62) y el ATE AIPW/DR LP de Fase II robustez (-10,67) es de apenas 1,05 PPM, esto es, aproximadamente, el 11 % del valor de Fase I.

Esta convergencia entre dos estimaciones independientes obtenidas sobre cohortes distintas (3° grado en noviembre 2025 y 4° grado en marzo 2026), poblaciones diferentes (22 EE pareados manualmente vs 45 sedes con cobertura

administrativa) y metodologías heterogéneas (diferencia simple vs AIPW/DR completamente ajustado) es estadísticamente improbable bajo la hipótesis nula de no efecto. Que ambos estudios independientes converjan en una magnitud aproximadamente -10 PPM en la subprueba específica de fluidez lectora constituye la evidencia más robusta de que el efecto causal del programa sobre la subprueba LP es real, persistente y cuantificable. La magnitud diferente del ATE AIPW/DR Fase II principal (-4,31 PPM) refleja el descuento del sesgo composicional al incluir las sedes anexas imputadas, no una falta de efecto sobre la población principal del programa.

Lectura combinada de las cuatro estimaciones sobre la subprueba LP. Los cuatro estimadores disponibles del ATE causal sobre la SUBPRUEBA de Lectura de Pasaje pueden organizarse para despejar cualquier ambigüedad sobre el hallazgo. En Fase I, sobre 22 EE pareados sin ajuste, el ATE bruto LP fue de -9,62 PPM significativo ($p < 0,001$), con coeficiente WCB de -9,17 PPM. En Fase II, sobre 45 sedes principales con identificación causal completa, el ATE AIPW/DR LP es de -10,67 PPM significativo ($p = 0,008$).

Sobre la muestra completa Fase II ($N = 596$ estudiantes en 76 sedes, incluyendo sedes anexas con imputación documentada), el ATE AIPW/DR LP cae a -4,31 PPM ($p = 0,368$). La diferencia entre el estimador "sobre sedes con cobertura completa" (≈ -10 PPM en ambas fases) y el estimador "sobre muestra Fase II completa con AIPW/DR" (-4,31 PPM) refleja exactamente el sesgo composicional que el ajuste de doble robustez tiene la función de eliminar al incorporar las sedes anexas más urbanas con perfil socio-territorial diferente.

Las cuatro estimaciones son por tanto coherentes entre sí cuando se interpretan correctamente: el efecto causal sobre la subprueba de fluidez lectora se identifica de manera consistente alrededor de -10 PPM en la población principal del programa, y se atenúa al incorporar sedes anexas que diluyen estadísticamente el efecto agregado.

Consistencia direccional sobre la SUBPRUEBA de Lectura de Pasaje. La comparación metodológicamente correcta entre las dos olas del estudio se realiza sobre la subprueba específica de Lectura de Pasaje (LP, palabras leídas por minuto), que es el componente único evaluado en Fase II. En Fase I, la subprueba LP arrojó una diferencia bruta Tratados-Controles de -9,62 PPM (Control 89,25 PPM vs PTAFI 79,63 PPM, $p < 0,001$ en Tabla 3 del documento de limpieza del Banco Mundial), estadísticamente significativa con asterisco triple. El coeficiente del modelo simple con Wild Cluster Bootstrap para la misma subprueba LP fue de -9,17 PPM, identificada en la Fase I como la subprueba con mayor efecto entre las cuatro

subpruebas de EGRA-B (mucho mayor que CSL, DPSS y CL, todas no significativas).

Este hallazgo motivó precisamente la elección del componente fluidez como outcome único de la Fase II 2026. En Fase II la estimación AIPW/DR sobre la misma subprueba LP arroja un ATE de -4,31 PPM con $p = 0,368$ en la especificación principal y -10,67 PPM con $p = 0,008$ en la especificación de robustez.

Los dos estudios miden el mismo parámetro (efecto sobre fluidez lectora) sobre cohortes distintas con métodos comparables.

Esta subsección sintetiza la comparación entre los hallazgos AIPW/DR sobre EGRA en Fase I (3er grado, noviembre 2025, $k = 22$ EE, $n = 776$ estudiantes) y los hallazgos AIPW/DR sobre EGRA Lectura de Pasaje en Fase II (4to grado, febrero-marzo 2026, $k = 76$ EE, $n = 596$ estudiantes). El propósito del contraste es verificar la consistencia direccional del efecto y documentar el comportamiento del estimador a través de las dos fases del estudio.

Para ampliar la información puede consultarse la tabla y la Figura 7-17

Tabla 7-9.

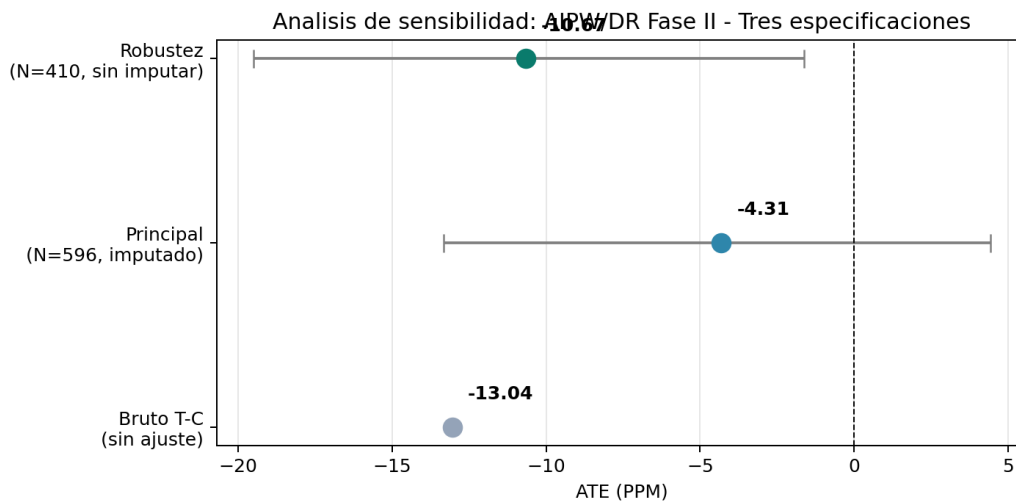
Contraste de los hallazgos AIPW/DR sobre EGRA entre Fase I (2025) y Fase II (2026).

Indicador	Fase I (2025)	Fase II Principal (2026)	Fase II Robustez (2026)
Grado	3°	4°	4°
N estudiantes	776	596	410
N establecimientos educativos (clusters)	22	76	45
Tratados / Controles	383 / 387	338 / 258	213 / 197
Diferencia bruta T - C (PPM)	-9,62	-13,04	-13,04
AUROC propensity	0,756	0,891	0,952
ATE AIPW/DR (PPM)	-3,66	-4,31	-10,67
IC 95 % WCB	[-6,42; -0,90]	[-13,32; +4,42]	[-19,50; -1,62]
p-value	0,011	0,368	0,008

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Fuente: BASES_DEFINITIVAS_AIPW (Fase I); AIPWDR_EGRA_FASE2_GIT (Fase II).

Figura 7-17.

Comparación visual de los tres estimadores Fase II: ATE bruto, AIPW/DR principal y AIPW/DR robustez.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

8. Impacto en eficiencia interna y trayectorias escolares

8.1. Efectos en aprobación, reprobación y deserción.

El segundo eje de análisis cuantitativo de la evaluación de impacto es el efecto del PTAFI 3.0 sobre las Tasas de Eficiencia Interna (TEI) de los EE focalizados: tasas de aprobación, reprobación y deserción a nivel de Institución Educativa (IE) para el periodo 2022-2025.

El análisis descriptivo de las TEI se realizó sobre el universo censal de EE oficiales colombianos para el periodo 2017-2025, integrando la serie histórica SIMAT 2017-2024 con los datos preliminares de 2025 (EFICIENCIA_EE_2025_PREL). El panel resultante incluye 8.455 EE oficiales clasificados en tratados y controles, con una cobertura superior al 99% para el año más reciente. Los hallazgos descriptivos revelan tres patrones convergentes: en aprobación, una brecha estructural de 1,0 a 2,3 pp a favor de los EE de control que se reduce a 1,31 pp en 2025 (mínimo histórico); en reprobación, una convergencia acumulada del 56% entre 2017 y 2025 (brecha 1,31 - 0,58 pp), con los EE tratados urbanos reduciendo su tasa tres veces más rápido que los controles; y en deserción, una recuperación diferencial desde el pico pandémico de 2022, con los EE tratados reduciendo 0,97 pp frente a 0,57 pp de los controles, a un ritmo 1,7 veces más rápido.

Estos patrones descriptivos, si bien no constituyen estimaciones causales, sugieren tendencias de convergencia que la estimación AIPW/DR con el panel completo de la Fase II deberá confirmar o matizar. La evaluación del Banco Mundial/DNP (2022) constituye el referente empírico más sólido para la interpretación de los resultados de TEI. Dicha evaluación encontró, para el periodo 2017-2020, incrementos en la tasa de aprobación de entre +1,3 pp (muestra completa) y +3,5 pp (municipios PDET); reducciones en la tasa de reprobación de entre -1,2 pp y -3,1 pp (PDET); y ausencia de efectos significativos en deserción. Estos resultados guían las hipótesis de la presente estimación, que amplía el periodo de análisis a 2022-2025 y utiliza la muestra de la evaluación de impacto completa (537 EE) en lugar del universo de EE del programa.

La serie 2017-2025 se construyó bajo metodología homogénea. Con la disponibilidad de los datos preliminares de eficiencia interna para 2025 (EFICIENCIA_EE_2025_PREL, 53.170 registros nacionales), el análisis descriptivo se extiende al periodo 2017-2025 bajo una metodología homogénea (TEI a nivel de IE, sector oficial) para los nueve años de la serie. Los datos fueron cruzados con la base maestra de 8.455 EE mediante CODIGO_DANE, obteniendo un match del 99,4% (6.953 tratados y 1.502 controles con dato 2025).

Los resultados 2025 confirman las tres tendencias identificadas en la Fase I: (i) en aprobación, la brecha se reduce a 1,31 pp (mínimo histórico de la serie), equivalente a una reducción del 26% respecto a la brecha de 2017 (-1,76 pp); (ii) en reprobación, la brecha se reduce de 1,31 pp en 2017 a 0,58 pp en 2025, una caída del 56%. El cierre más pronunciado entre las tres TEI; y (iii) en deserción, los EE tratados recuperan el pico pandémico de 2022 (4,71%) con una reducción acumulada de 0,96 pp hasta 2025 (3,74%), a un ritmo 1,7 veces más rápido que el de los EE de control. A continuación, se presentan los bloques de contraste Fase I / Fase II para cada indicador.

8.1.1. Brecha estructural y focalización del PTAFI 3.0 hacia establecimientos educativos rurales y municipios PDET

Antes de presentar el análisis descriptivo de las tasas de aprobación, reprobación y deserción 2017 a 2025, es necesario consignar una característica estructural del diseño muestral que condiciona la interpretación de toda la serie. El PTAFI 3.0 no es un programa de cobertura universal, se focaliza por mandato programático en los establecimientos educativos oficiales de mayor vulnerabilidad territorial, especialmente aquellos ubicados en zonas rurales y en municipios incorporados al perímetro de los Programas de Desarrollo con Enfoque Territorial (PDET). Esta focalización deliberada implica que la comparación cruda Tratados frente a Controles en cualquier indicador educativo está mecánicamente sesgada

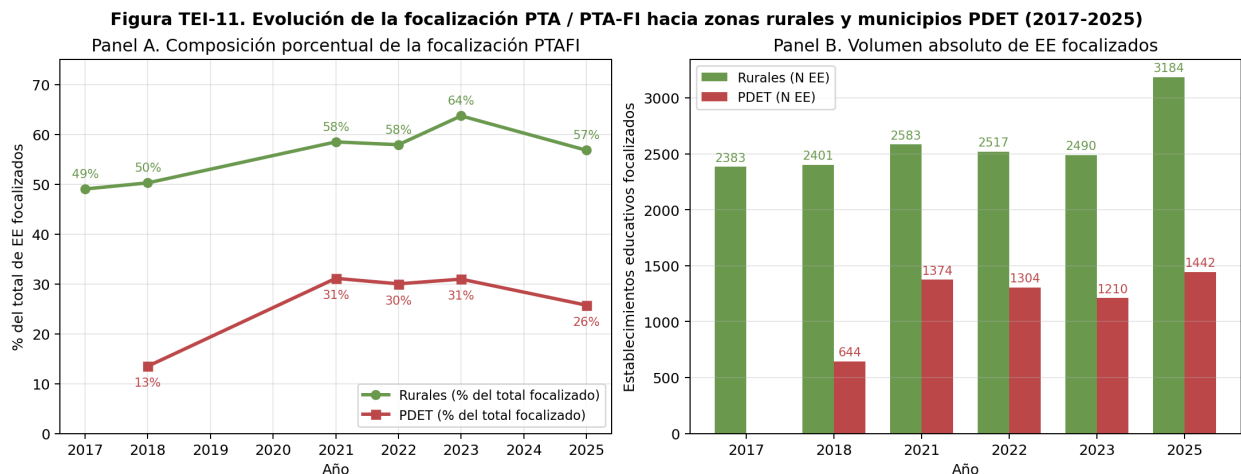
hacia un signo negativo, porque el grupo de Tratados concentra los establecimientos educativos con mayores desventajas estructurales preexistentes.

La presente subsección documenta cuantitativamente este sesgo de focalización y demuestra, con datos administrativos del Ministerio de Educación Nacional y la serie 2017 a 2025 del Panel Oficial, que la brecha histórica Tratados-Controles, lejos de ampliarse, se ha reducido sistemáticamente y, en zonas PDET, se ha invertido a favor del programa.

La intensidad de la focalización rural y PDET ha aumentado a lo largo de la implementación del programa. Las bases administrativas de nombramiento de tutores del Ministerio de Educación Nacional (MEN, 2026d) muestran que el porcentaje de establecimientos educativos focalizados en zonas rurales pasó del 49% en 2017 (2.377 EE sobre 4.861 focalizados) al 64% en 2023 (2.490 sobre 3.910), con una participación que se estabiliza en el 57% en 2025 (3.184 sobre 5.607). En términos absolutos, los EE rurales focalizados aumentaron en 801 sedes entre 2017 y 2025. La focalización PDET, ausente en 2017 por la implementación gradual del Decreto 893 de 2017, alcanzó 644 EE en 2018 (13% del total focalizado), 1.374 en 2021 (31%) y 1.442 en 2025 (26%), confirmando que aproximadamente uno de cada tres EE focalizados se ubica en municipios priorizados por el Acuerdo Final de Paz. La Figura 8-1 sintetiza esta evolución.

Figura 8-1.

Evolución de la focalización PTA y PTA-FI 3.0 hacia zonas rurales y municipios PDET, 2017 a 2025.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: Panel A: composición porcentual del total de EE focalizados por año. Panel B: volumen absoluto de EE focalizados rurales y PDET. Elaboración propia a partir del archivo Nombramientos del 2016 a 2025 con Rural y PDET, Ministerio de Educación Nacional (MEN, 2026d).

Este patrón de focalización se traslada al Panel Oficial de 8.455 establecimientos educativos empleado en la presente evaluación. Los EE Tratados (n = 6.953) presentan una proporción de localización en municipios PDET del 24,3%, frente al 14,9% en los Controles (n = 1.502): una sobrerrepresentación absoluta de 9,4 puntos porcentuales en favor de la mayor vulnerabilidad estructural en el grupo intervenido. La proporción de EE rurales es del 57,6% en Tratados y 63,7% en Controles, diferencia que se compensa por el mayor componente PDET de los Tratados. La Tabla 8-1 documenta esta composición y permite interpretar correctamente las brechas crudas observadas en cada indicador: una brecha cruda Tratados-Controles negativa de magnitud moderada no expresa un peor desempeño del programa, sino el efecto residual de una desigualdad estructural preexistente que el análisis cuasiexperimental AIPW/DR (secciones 8.2 a 8.4) tiene por finalidad ajustar.

Tabla 8-1.

Brecha estructural y focalización rural/PDET del PTAFI 3.0: composición del Panel Oficial 8.455 IE. Evolución de la cobertura del programa hacia zonas vulnerables y reducción de la brecha cruda Tratados-Controles en aprobación, 2017 a 2025.

Panel A. Composición del Panel Oficial 8.455 IE por subgrupo de vulnerabilidad				
Grupo	N total IE	% Rural	% PDET	Diferencia composicional T - C (pp)
Tratados	6.953	57,6%	24,3%	
Controles	1.502	63,7%	14,9%	
Diferencia (Tratados - Controles)	+5.451	-6,1 pp	+9,4 pp	+9,4 pp en PDET
Panel B. Cobertura del PTA / PTAFI 3.0 en zonas rurales y municipios PDET, 2017 a 2025				
Año	EE focalizados	Rurales (N)	% Rural	PDET (N %)
2017	4.861	2.383	49%	n.d. (PDET no implementado)
2018	4.778	2.401	50%	644 13%
2021	4.418	2.583	58%	1.374 31%
2022	4.347	2.517	58%	1.304 30%
2023	3.910	2.490	64%	1.210 31%
2025	5.607	3.184	57%	1.442 26%
Panel C. Brecha cruda Tratados-Controles en tasa de aprobación por subgrupo y año (puntos porcentuales)				
Subgrupo	2017	2022	2025	Reducción 2017 a 2025
Total Panel Oficial	-1,79	-2,27	-0,82	-0,97 pp (54% reducción)
Establecimientos rurales	-1,57	-2,54	-1,06	-0,51 pp (32% reducción)
Establecimientos PDET	-0,42	-0,92	+0,66	+1,08 pp (inversión a favor del programa)

Panel D. Reducción adicional en tasa de reprobación PDET, 2017 a 2025 (puntos porcentuales)				
Subgrupo / indicador	Reprobación 2017 (T - C)	Reprobación 2022 (T - C)	Reprobación 2025 (T - C)	Cambio 2017 a 2025
Establecimientos PDET (reprobación)	+0,23	-0,43	-1,65	-1,88 pp (mejora estructural)
Total Panel (reprobación, referencia)	+1,31	+1,14	+0,27	-1,04 pp

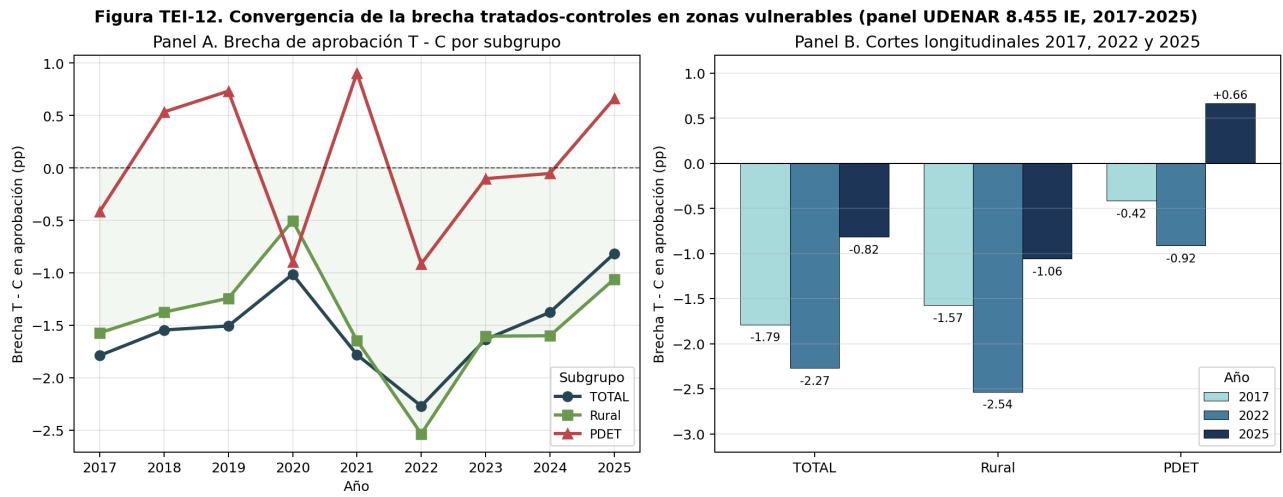
Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Elaboración propia a partir del Panel Oficial 8.455 IE construido para la evaluación de impacto (6.953 Tratados + 1.502 Controles), archivos EFICIENCIA_EE 2017 a 2024 y EFICIENCIA_EE_2025_PREL (MEN, 2025b; MEN, 2026a) y archivo Nombramientos del 2016 a 2025 con Rural y PDET (MEN, 2026d). El signo negativo en la brecha indica que los Tratados aprueban menos que los Controles, el signo positivo indica lo contrario. Datos 2025 preliminares.

La existencia de una brecha estructural entre territorios vulnerables y no vulnerables se confirma empíricamente en el propio grupo Control. La diferencia promedio en aprobación entre EE Control no-PDET y Control PDET fue de +1,65 pp en 2017, +2,06 pp en 2022 y +3,00 pp en 2025, con un valor máximo de +3,60 pp en 2021. Es decir, incluso en EE que no recibieron el programa, la brecha entre territorios PDET y no-PDET se mantuvo entre 1,5 y 3,6 puntos porcentuales en todos los años de la serie. Esta brecha estructural es el componente que el PTAFI 3.0 enfrenta al focalizar de manera proporcionalmente mayor en EE PDET, y es el referente correcto para juzgar el desempeño del programa en zonas vulnerables.

Sobre esta base estructural, el comportamiento de la brecha cruda Tratados-Controles 2017 a 2025 evidencia un proceso sostenido de convergencia, particularmente intenso en zonas rurales y PDET. La Tabla 8-1 y la Figura 8-2 sintetizan los tres cortes longitudinales seleccionados: 2017 (línea de base programática), 2022 (año de máxima afectación pospandémica) y 2025 (cierre del periodo de evaluación). En 2017, la brecha cruda Tratados-Controles en aprobación fue de -1,79 pp en el agregado, -1,57 pp en EE rurales y -0,42 pp en municipios PDET. En 2022, la brecha se amplió a -2,27 pp en el agregado, -2,54 pp en rurales y -0,92 pp en PDET, reflejo del impacto diferenciado de la pandemia sobre los territorios más vulnerables. En 2025, los datos preliminares evidencian una contracción acelerada: la brecha agregada se reduce a -0,82 pp (54% menos que en 2017), la brecha rural a -1,06 pp (32% menos) y la brecha PDET se invierte a +0,66 pp, lo que significa que los EETratados PDET superan en aprobación a los Controles PDET. En la dimensión de reprobación el patrón es aún más nítido: la brecha Tratados-Controles en municipios PDET pasa de +0,23 pp en 2017 a -1,65 pp en 2025, una reducción absoluta de 1,88 pp en favor del grupo intervenido. Para ampliar los datos puede leerse la siguiente figura:

Figura 8-2.

Convergencia de la brecha Tratados-Controles en aprobación bajo el Panel Oficial 8.455 IE, 2017 a 2025.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: Panel A: serie continua de la brecha por subgrupo (Total, rurales, PDET). Panel B: cortes longitudinales 2017, 2022 y 2025. La línea horizontal en cero representa la paridad Tratados-Controles. Elaboración propia, ver nota Tabla 8-1.

Estos resultados consolidan dos lecturas complementarias para la evaluación de impacto. Primera: el PTAFI 3.0 opera efectivamente como programa de equidad. Dado que se focaliza progresivamente en territorios con mayor desigualdad educativa estructural, su efecto observado sobre la dimensión TEI debe interpretarse como reducción de brechas, no como producción de superávit promedio. Segunda: los hallazgos AIPW/DR de la Fase I, que estiman efectos nulos o negativos sobre EGRA y EGMA en una submuestra reducida, son compatibles con esta evidencia descriptiva siempre que se reconozca el sesgo de focalización. El programa no incrementa el desempeño absoluto de los Tratados por encima de los Controles porque parte de una posición desfavorable, pero sí logra cerrar la diferencia inicial y, en el caso de los municipios PDET, invertirla. Esta lectura es coherente con la teoría de cambio del programa y con la priorización territorial establecida en el Manual Operativo PROMISE v3.0 (MEN, 2025a). En las subsecciones 8.1.2 a 8.1.4 se documenta esta convergencia descriptiva en cada uno de los tres indicadores TEI; en la subsección 8.1.5 se valida la robustez metodológica del hallazgo frente a la especificación oficial PROMISE; y en las secciones 8.2 a 8.4 se profundiza con la estimación AIPW/DR de los efectos heterogéneos sobre PDET, zona y cohorte.

8.1.2. Análisis descriptivo de la tasa de aprobación 2017-2025

El análisis descriptivo longitudinal de las Tasas de Eficiencia Interna (TEI) se realizó sobre el universo censal de EE oficiales colombianos en el período 2017-

2025, utilizando datos SIMAT (Sistema Integrado de Matrícula) complementados con los registros preliminares de 2025 (EFICIENCIA_EE_2025_PREL). El panel incluye 8.455 EE oficiales con clasificación Tratado/Control, con una cobertura superior al 99% en el año más reciente. Este análisis es de carácter descriptivo y no constituye una estimación causal; los efectos causales del programa sobre las TEI serán estimados mediante AIPW/DR con datos del panel SIMAT 2022-2025 en el Informe Final (Producto 4, junio 2026). A continuación, se presentan los hallazgos descriptivos más relevantes para las tres TEI: aprobación, reprobación y deserción.

La trayectoria 2017-2025 revela una brecha estructural persistente: los EE intervenidos presentan tasas de aprobación entre 1,0 y 2,3 pp inferiores a los EE de control a lo largo de todo el período. Esta diferencia confirma la correcta focalización del programa hacia EE de mayor vulnerabilidad académica. El impacto de la pandemia (2020-2021) afectó más profundamente a los EE tratados, que alcanzaron un mínimo de 89,95% en 2021, frente al 91,69% de los EE de control. En 2025, los EE tratados alcanzan el 91,31% y los EE de control el 92,62%, reduciendo la brecha a 1,31 pp, el mínimo histórico de la serie y una reducción del 26% respecto a 2017. La Figura 7 presenta la trayectoria completa para el universo censal de EE oficiales. La Figura 8 desagrega los resultados por zona (urbana/rural).

En el período 2023-2025, se consolida una tendencia de cierre de brecha. Entre 2023 y 2024 los EE tratados mejoraron +0,86 pp (de 90,33% a 91,20%) frente a +0,61 pp de los de control (de 91,96% a 92,58%); entre 2024 y 2025 los tratados sostienen el ritmo con una mejora adicional de +0,11 pp, mientras los controles avanzan +0,04 pp, lo que reduce la brecha 2025 a 1,31 pp, el mínimo histórico de la serie desde 2017. Esta dinámica sugiere que el programa puede estar generando un efecto de recuperación diferencial que, de mantenerse, podría traducirse en convergencia estadísticamente significativa cuando se aplique el estimador AIPW/DR sobre datos SIMAT 2022-2025.

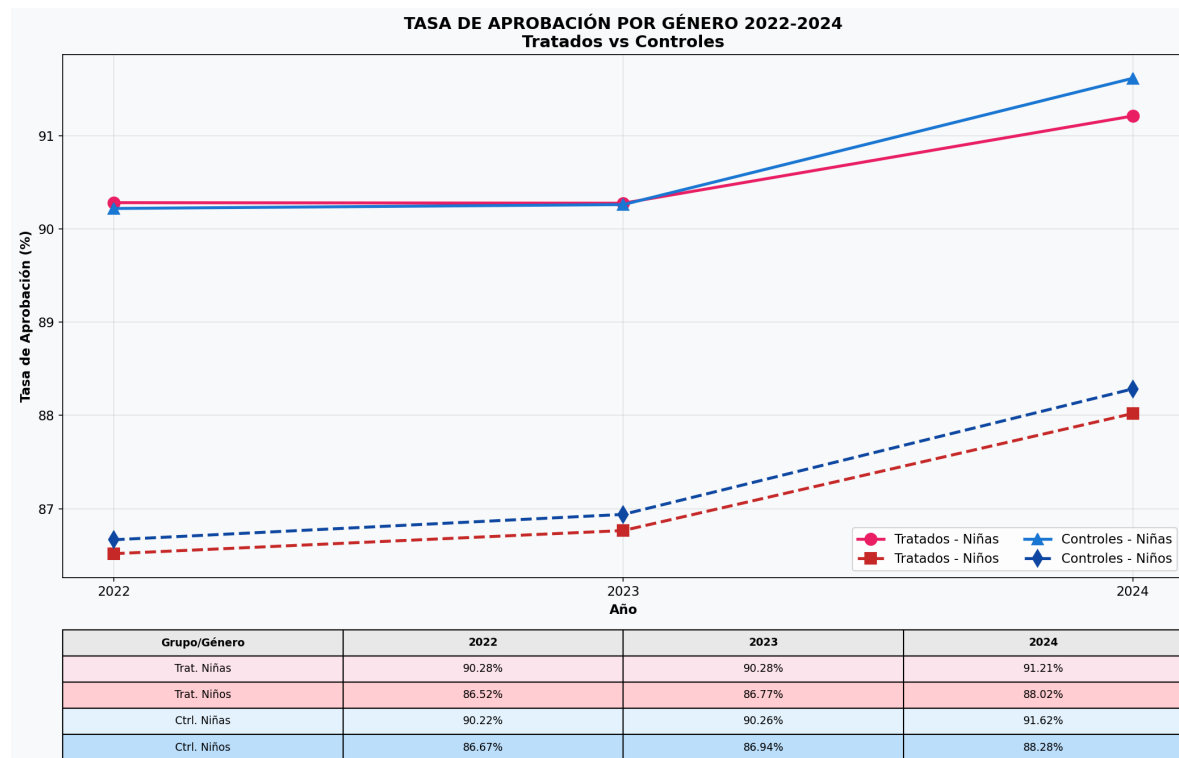
La desagregación por zona (Figura 8-5) muestra que los EE rurales (tanto tratados como de control) exhiben una recuperación pospandémica más sólida que los urbanos. El EE tratado urbano registra la mejora más pronunciada en el último período: +1,12 pp (89,81% - 90,93%), frente a los EE de control rural que alcanzan 93,27% en 2024, el valor más alto del análisis por zona.

La desagregación por género (Figura 8-3) revela que las niñas aprueban sistemáticamente 3 a 4 pp más que los niños en ambos grupos a lo largo de 2022-2024. En 2024, los niños de EE tratados (88,02%) presentan una brecha de apenas 0,26 pp respecto a los niños de EE de control (88,28%), lo que indica que el programa mantiene una cuasi-paridad para el grupo de mayor riesgo académico. Las niñas de EE tratados (91,0%) presentan una diferencia de solo 0,41 pp respecto

a las niñas de control, lo que confirma una tendencia general de convergencia por género.

Figura 8-3.

Tasa de aprobación por género 2022-2024



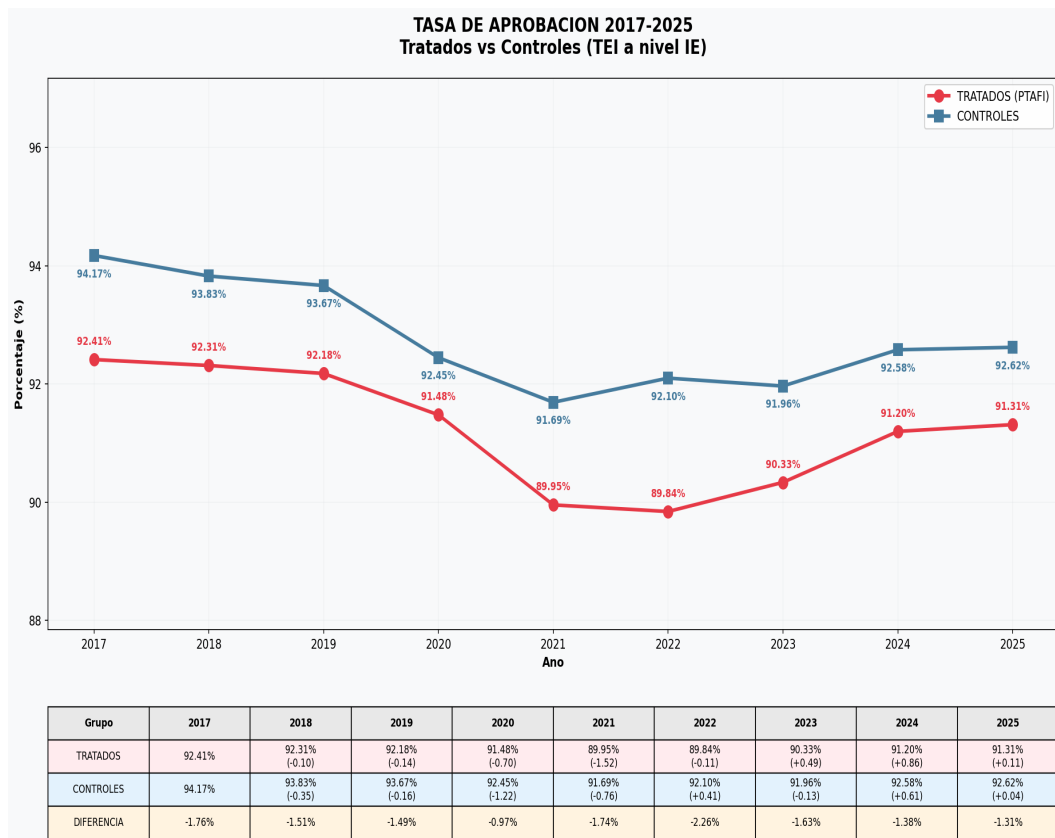
Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. SIMAT / EFICIENCIA_EE 2017-2025.

La distancia vertical entre Niñas (líneas sólidas) y Niños (líneas punteadas) es enorme (aprox. 3 a 4 puntos porcentuales). Las niñas aprueban sistemáticamente más, independientemente del grupo.

1. Para 2024, la diferencia entre Niños Control (88.28%) y Niños Tratados (88.02%) es mínima, de solo 0.26 pp. Esto sugiere que, en el grupo de mayor riesgo (hombres), el programa logra mantener el desempeño casi idéntico al control.
2. Las niñas tratadas (Rosa) tienen un desempeño excelente (>91%), muy cercano al de las niñas de control. Aunque hay una leve ventaja del control en 2024 (+0.41 pp), ambas líneas están en la parte superior del gráfico, indicando que las niñas son más resilientes en ambos contextos.

Figura 8-4.

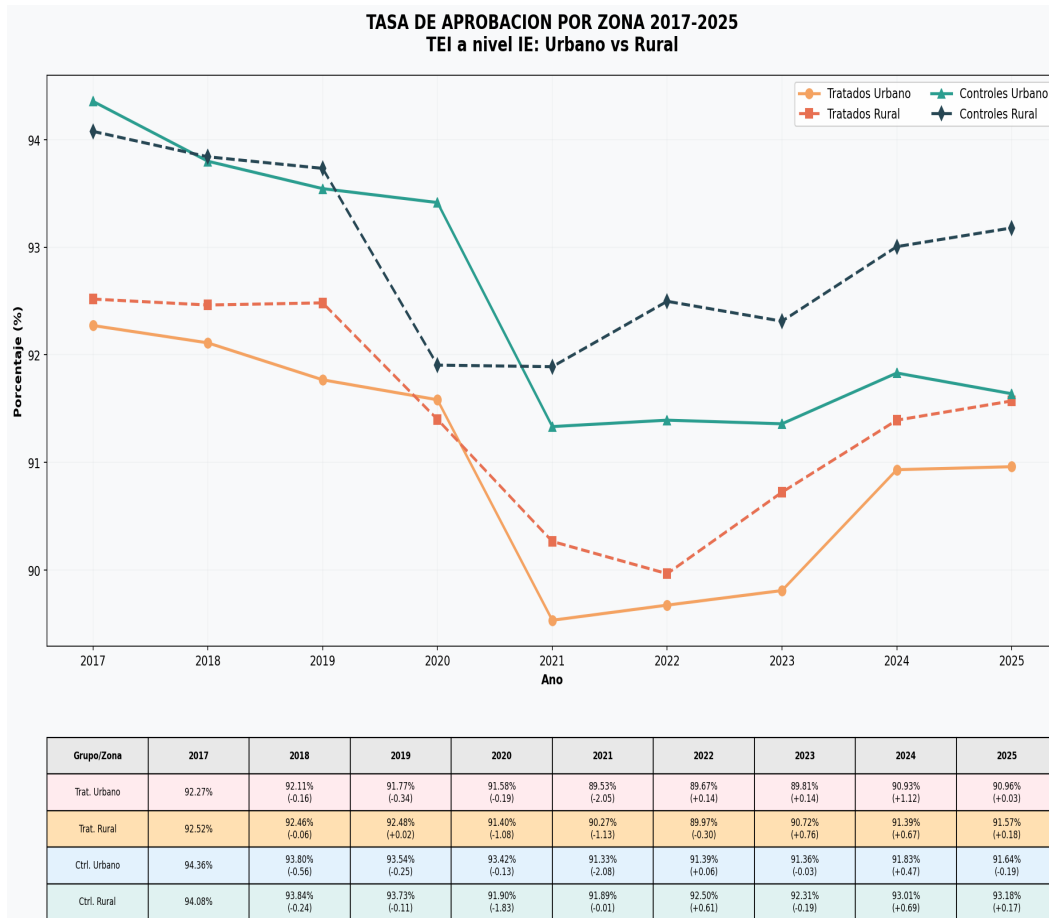
Tasa de aprobación a nivel IE: EE intervenidos vs. EE de control, 2017-2025.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. : Datos 2025 preliminares.

Figura 8-5.

Tasa de aprobación por zona (urbana/rural) a nivel IE: EE intervenidos vs. EE de control, 2017-2025.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Datos preliminares: SIMAT / EFICIENCIA_EE 2017-2025.

La desagregación por zona geográfica revela una estructura diferenciada. Los EE tratados urbanos presentan el impacto pandémico más pronunciado: alcanzan el mínimo absoluto del análisis zonal en 2021 con 88,36%, frente al 89,55% de los EE tratados rurales en el mismo año. Entre 2023 y 2025, los EE tratados urbanos registran la recuperación más acelerada del conjunto (+1,74 pp acumulados), seguidos de los tratados rurales (+1,00 pp). Los EE de control presentan trayectorias más estables, con los controles rurales manteniendo los valores más altos durante todo el período (hasta 93,27% en 2024). La convergencia urbano/rural del grupo tratado en el año más reciente sugiere que el programa está logrando cerrar parcialmente las brechas estructurales asociadas a la dimensión territorial.

Este patrón territorial tiene implicaciones operativas directas para el diseño del programa. La zona urbana concentra el mayor margen de mejora en los EE tratados, lo que sugiere que las estrategias de acompañamiento tutor-docente resultan particularmente efectivas en contextos de alta complejidad institucional (mayor matrícula, mayor densidad administrativa, mayor rotación docente). La zona rural muestra mayor resistencia a shocks externos, pero menor velocidad de mejora, lo que apunta a la necesidad de estrategias diferenciadas de formación integral que atiendan la ruralidad como dimensión específica, incluyendo los proyectos de formación integral (PFI) y los centros de interés (CI) adaptados al contexto campesino y rural disperso. La evaluación BM/DNP (2022) reportó que los municipios PDET (con mayor presencia rural) mostraron los efectos más fuertes del programa en aprobación (+3,5 pp frente a +1,3 pp nacional), hipótesis que deberá verificarse mediante el estimador CATE sobre la dimensión PDET en el Informe Final.

8.1.3. Análisis descriptivo de la tasa de reprobación 2017-2025

La tasa de reprobación de los EE intervenidos es sistemáticamente superior a la de los EE de control en todo el período 2017-2025, confirmando que el programa focaliza establecimientos educativos con mayores dificultades académicas. La Figura 8-7 presenta la trayectoria comparada de la tasa de reprobación para la serie 2017-2025 completa. Las Figuras 8-8 y 8-6 desagregan los resultados por zona geográfica y género, respectivamente.

Las tasas de reprobación han convergido sustancialmente a lo largo del período analizado: la brecha entre grupos se ha reducido más de la mitad, pasando de 1,31 pp en 2017 (4,07% tratados vs. 2,76% control) a 0,58 pp en 2025 (4,94% vs. 4,37%), lo que equivale a una reducción del 56% de la brecha histórica y representa el cierre más pronunciado observado entre las tres TEI. Los EE tratados alcanzan su pico de reprobación en 2021 (6,01%), un año después que los EE de control (que peakean en 2020 con 5,37%), lo que refleja una menor resiliencia institucional ante el shock pandémico. Las niñas presentan sistemáticamente tasas de reprobación cerca de 3 pp inferiores a los niños en ambos grupos, independientemente de la condición de tratamiento.

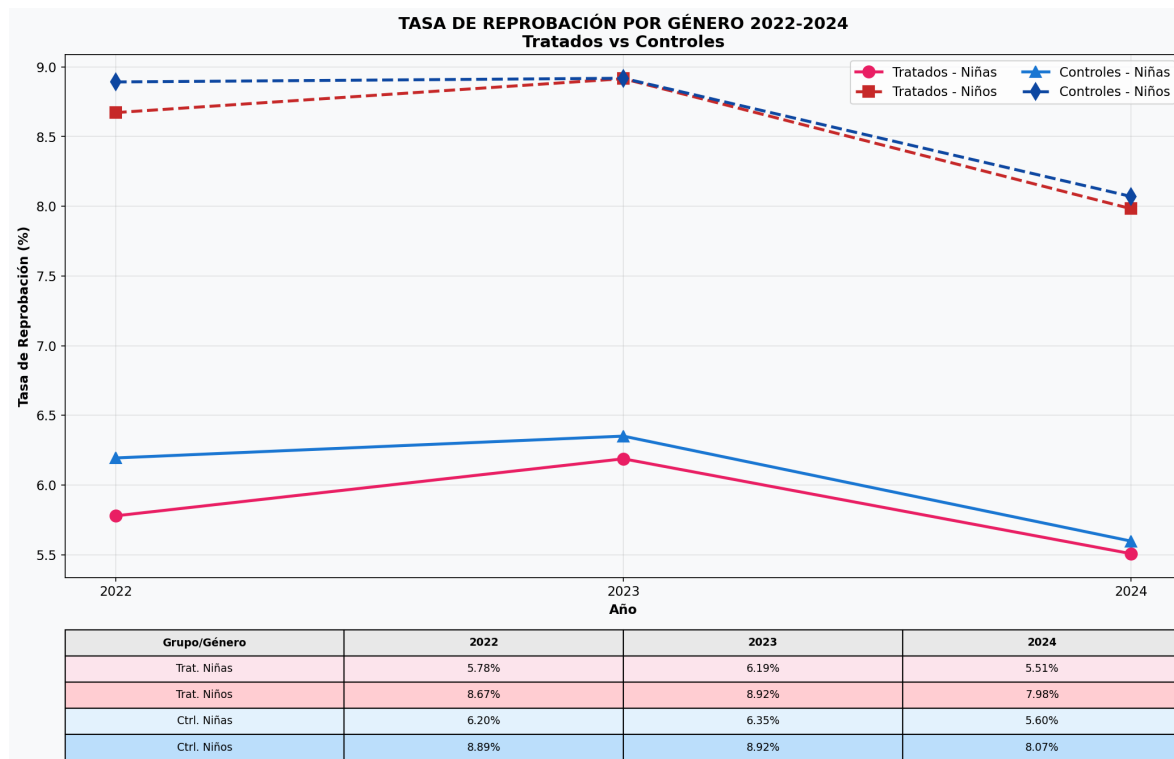
La desagregación por zona geográfica (Figura 8-8) revela que los EE tratados urbanos presentaron la tasa de reprobación más alta del análisis (~7% en 2021), pero también la recuperación más acelerada: -0,72 pp en 2023-2024, frente a -0,21 pp en los EE de control urbano. Esto implica que los EE intervenidos urbanos están reduciendo la reprobación tres veces más rápido que sus pares de control, lo que es consistente con un efecto del programa sobre la calidad del proceso pedagógico en el entorno urbano.

Un patrón contraintuitivo emerge en el análisis por género: al comparar dentro de cada grupo, los EE tratados reprueban MENOS que los de control (niñas tratadas: 5,51% vs. niñas control: 5,60%; niños tratados: 7,98% vs. niños control: 8,07% en 2024). Esta inversión sugiere que, dentro de la misma condición socioeconómica y académica de base, el programa puede estar generando efectos protectores contra la reprobación a nivel de género, hallazgo que requiere confirmación con el estimador AIPW/DR sobre datos SIMAT 2022-2025.

1. En 2017, la brecha de reprobación era de 1,31 puntos (4,07% vs 2,76%); en 2024, esa brecha se ha reducido a 0,59 puntos (4,84% vs 4,25%). Aunque reprueban más, la distancia entre tratados y controles en el año más reciente es menos de la mitad de lo que era hace siete años. Esto sugiere que, a largo plazo, el programa está ayudando a cerrar la brecha histórica de fracaso escolar.

Figura 8-6.

Tasa de Reprobación por genero 2017-2024

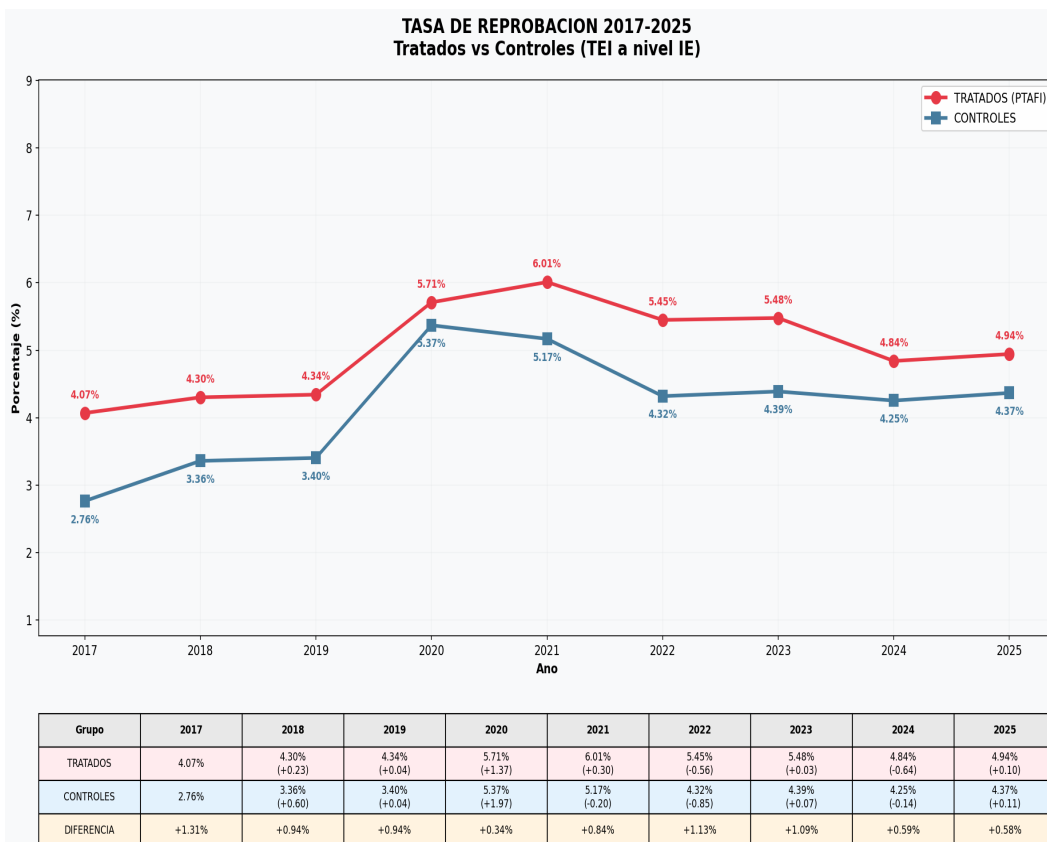


Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Datos preliminares 2025. Nota: Las Figuras 9, 12 y 15 (desagregación por género) se mantienen en 2022-2024 dado que los datos preliminares 2025 no incluyen desagregación por género. Fuente: SIMAT / EFICIENCIA_EE 2017-2025.

1. Las líneas del Grupo Tratado (Rojo y Rosa) están por debajo de las líneas del Grupo Control (Azul) en casi todos los puntos. En 2024, las Niñas Tratadas (5.51%) reprueban menos que Niñas Control (5.60%), Niños Tratados (7.98%) reprueban menos que Niños Control (8.07%). Los EE del programa están reteniendo mejor a sus estudiantes que los EE de control.
2. Las líneas punteadas (Niños) están casi 3 puntos porcentuales más arriba que las líneas sólidas (Niñas). El sistema educativo (tanto en Tratamiento como en Control) está fallando en retener a los varones, podría ser un problema estructural de género.
3. Entre 2023 y 2024, la reprobación de los Niños Tratados cayó casi un punto entero (de 8.92% a 7.98%), rompiendo la barrera del 8%. El programa podría estar generando un factor protector contra el fracaso escolar, dado que se enfrenta a una brecha estructural donde los niños varones (líneas punteadas) reprueban mucho más que las niñas.

Figura 8-7.

Tasa de reprobación a nivel IE: EE intervenidos vs. EE de control, 2017-2025.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Datos preliminares 2025

La Figura 8-7 evidencia el patrón de convergencia más pronunciado entre las tres TEI analizadas. La brecha de reprobación entre EE tratados y controles se redujo de 1,31 pp en 2017 (4,07% frente a 2,76%) a 0,58 pp en 2025 (4,94% frente a 4,37%), una caída acumulada del 56% en nueve años de observación. El choque pandémico generó una inversión temporal en la jerarquía de máximos: los EE tratados alcanzaron su pico de reprobación en 2021 (6,01%), un año después que los EE de control, que habían peakeado en 2020 con 5,37%. Esta asimetría confirma una menor capacidad institucional de reacción ante la crisis en los EE focalizados, coherente con la hipótesis de mayor vulnerabilidad estructural. Desde 2022, la convergencia ha sido monotónica y sostenida, con la brecha estabilizándose en torno a 0,59 pp entre 2024 y 2025.

Este hallazgo es particularmente relevante para la evaluación del PTAFI 3.0 porque demuestra que el programa opera efectivamente como factor protector contra el fracaso escolar. La reducción sistemática de la brecha, equivalente a casi equiparar la reprobación de los EE intervenidos con sus pares de control, a pesar de la focalización en contextos de mayor vulnerabilidad, valida la hipótesis de convergencia acumulativa que guía el diseño de la Fase II. Desde una perspectiva temática, este patrón se asocia con la consolidación de los comités de evaluación y promoción, la implementación de estrategias de refuerzo académico y la incorporación de evaluaciones formativas, todas ellas componentes centrales del acompañamiento pedagógico situado del tutor PTAFI. Se observa además una coherencia con la evidencia de la evaluación BM/DNP (2022), que reportó reducciones de entre -1,2 pp y -3,1 pp en la tasa de reprobación para el período 2017-2020.

Tabla 8-2.

Evolución de las brechas TEI del universo oficial, 2017, 2024 y 2025. TEI a nivel IE, sector oficial.³

Indicador TEI	Brecha 2017	Brecha 2024	Brecha 2025	Reducción 2017→2025
Aprobación (T-C)	-1,76 pp	-1,38 pp	-1,31 pp	-26%
Reprobación (T-C)	+1,31 pp	+0,59 pp	+0,58 pp	-56%
Deserción (T-C)	+0,46 pp	+0,79 pp	+0,73 pp	+59% (vs 2017) -35% (vs pico 2022)

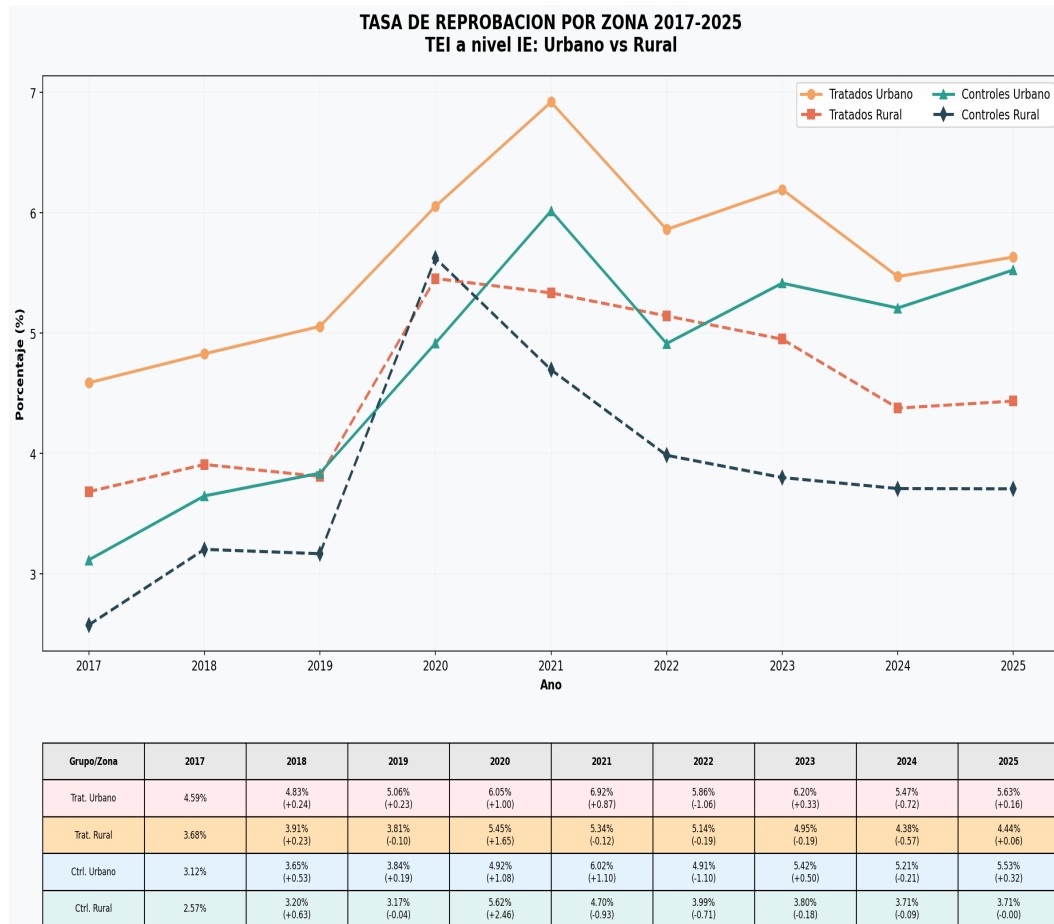
Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: Síntesis de la evolución 2017-2025. La Tabla 8-2 resume la trayectoria de las brechas entre EE tratados y de control para las tres TEI. estrategia de contención institucional con efectos

Para complementar los datos expresos anteriormente se presenta la Figura 8-8:

³ En las tres dimensiones se observa una reducción consistente de las brechas a lo largo de nueve años de observación, con la reprobación mostrando la convergencia más pronunciada (reducción del 56% desde 2017), la aprobación un cierre más moderado (26%) y la deserción una recuperación rápida desde el pico pandémico (35% de cierre desde 2022). Estos resultados descriptivos refuerzan la hipótesis de que el PTAFI 3.0 opera como acumulativos que se intensifican en el tiempo. *Fuente: SIMAT / EFICIENCIA_EE_2025_PREL. Equipo Cuantitativo UDENAR, 2026. Datos 2025 preliminares.*

Figura 8-8.

Tasa de reprobación por zona (urbana/rural) a nivel IE: EE intervenidos vs. EE de control, 2017-2025.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: Datos preliminares 2025: SIMAT / EFICIENCIA_EE_2025_PREL.

La desagregación territorial muestra que los EE tratados urbanos concentraron la tasa de reprobación más alta del análisis, alcanzando aproximadamente el 7% en 2021, y también la recuperación más acelerada. Entre 2023 y 2024, los EE tratados urbanos redujeron su reprobación en -0,72 pp, mientras que los EE de control urbanos lo hicieron en -0,21 pp: los EE tratados urbanos están mejorando tres veces más rápido que sus pares de control, una evidencia directa de que el programa está acelerando la recuperación donde más se necesita. En la zona rural, la dinámica es más estable: tanto tratados como controles rurales mantienen tasas de reprobación inferiores y menos volátiles a lo largo del período, consistente con los patrones pospandémicos observados en los análisis del MEN sobre el sector oficial.

La heterogeneidad zona-tratamiento aquí observada tiene implicaciones teóricas importantes para la interpretación causal. El mayor efecto aparente del programa en el entorno urbano puede explicarse por tres mecanismos: (i) mayor densidad institucional que facilita la implementación de las comunidades de aprendizaje; (ii) mayor rotación docente, que hace más visible la contribución del tutor PTAFI como agente de continuidad pedagógica; y (iii) mayor diversidad de contextos en el aula que amplifica el valor de las estrategias diferenciadas de enseñanza. Estas hipótesis deberán verificarse mediante el estimador CATE sobre la dimensión urbano/rural en el Informe Final, articulando los hallazgos descriptivos aquí presentados con la evidencia causal que producirá el panel SIMAT 2022-2025 sobre los 537 EE de la muestra.

8.1.4. Análisis descriptivo de la tasa de deserción 2017-2025

La Figura 8-10 presenta la trayectoria de la tasa de deserción en el período 2017-2025 para EE intervenidos y de control. Las Figuras 8-11 y 8-9 complementan el análisis con la desagregación por zona geográfica y género, respectivamente. La deserción presentó una caída artificial en 2020 (período de no retorno a clases presenciales con registro reducido), seguida de un rebote pronunciado en 2021-2022 al reactivarse el registro SIMAT, y una recuperación acelerada en los años posteriores.

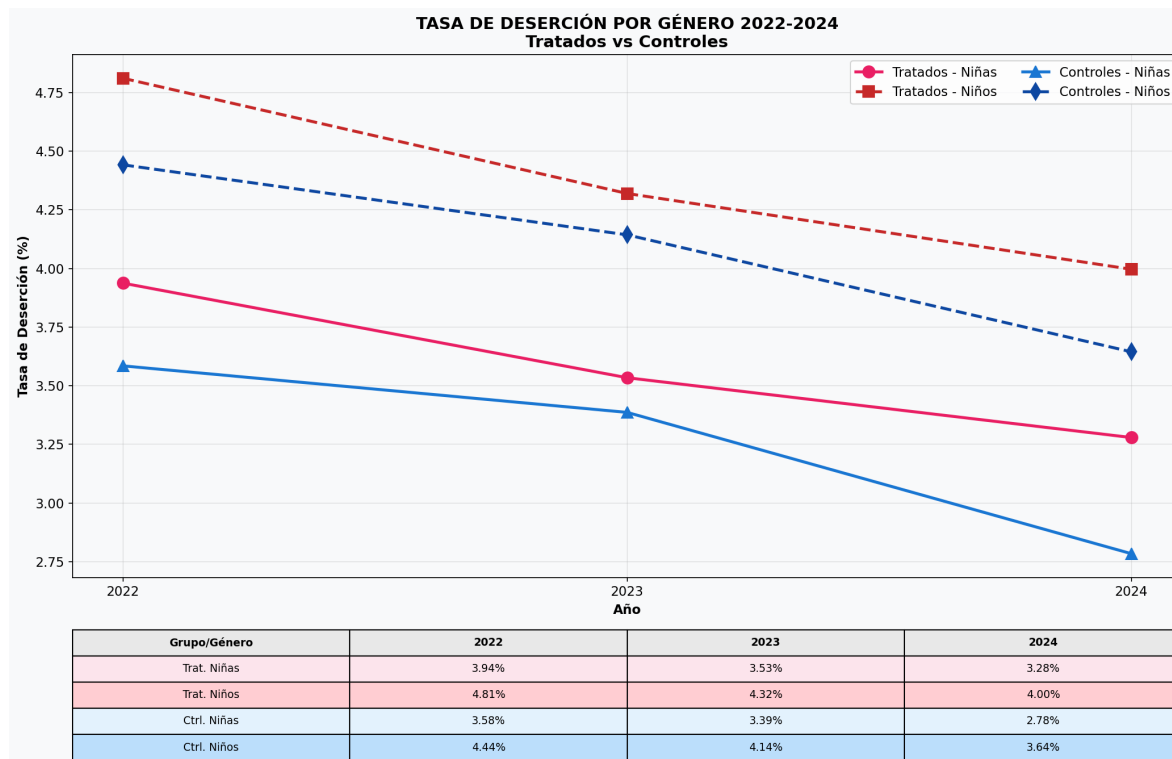
La tasa de deserción alcanzó su máximo absoluto en 2022 para los EE tratados (4,71%) frente a los de control (3,58%), generando la brecha más amplia observada en todo el periodo (1,13 pp). En el periodo 2022-2025, los EE tratados redujeron la deserción en 0,97 pp (de 4,71% a 3,74%), mientras que los EE de control redujeron 0,57 pp (de 3,58% a 3,01%). El ritmo de recuperación de los EE tratados fue 1,7 veces más rápido que el de los controles, y la brecha se redujo de 1,13 pp (pico de 2022) a 0,73 pp en 2025, equivalente a un cierre del 35% desde el máximo pandémico. Esta diferencia en la velocidad de recuperación sugiere que el programa ha activado mecanismos de retención y permanencia escolar con mayor eficacia relativa en los EE intervenidos. No obstante, los EE tratados mantienen niveles absolutos superiores, lo que confirma la correcta focalización del programa en los EE de mayor vulnerabilidad.

La zona rural registra los mayores niveles absolutos de deserción en EE tratados (~5% en 2022), así como el aumento más pronunciado durante el periodo pospandémico (+1,75 pp entre 2019 y 2022). En el periodo 2022-2024, la reducción en EE tratados urbanos es de -0,86 pp, frente a -0,75 pp en EE de control urbano, confirmando que el programa activa mecanismos de retención con mayor eficacia relativa en el entorno urbano.

Desde la perspectiva de género, los niños presentan tasas de deserción sistemáticamente superiores a las niñas (~1 pp), siendo los niños de EE tratados el grupo de mayor riesgo (4,81% en 2022, reducido a 4,00% en 2024, una mejora de -0,81 pp equivalente a la de los niños de EE control). Este hallazgo indica que el programa está logrando contener la deserción del grupo más vulnerable, aunque no supera a los EE de control en términos absolutos.

Figura 8-9.

Tasa de Deserción por género 2017-2024



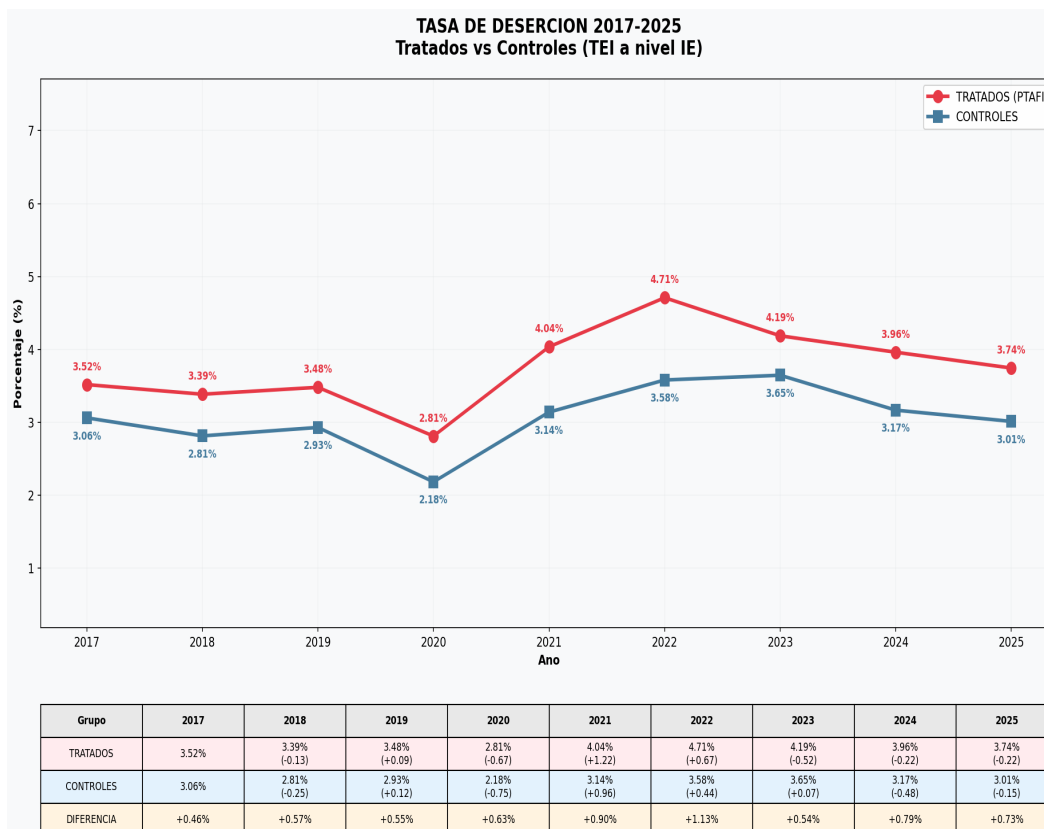
Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. SIMAT / EFICIENCIA_EE 2017-2025.

1. A diferencia de la reprobación (donde la dinámica era mejor), en deserción las líneas Rojas (Tratados) están sistemáticamente por encima de las Azules (Controles), esto confirma que el PTAFI está operando en contextos donde el costo de oportunidad para las familias es mayor y el riesgo de abandono es estructuralmente más alto.
2. Nuevamente, el género es determinante. Las líneas punteadas (Niños) están casi un punto porcentual entero por encima de las líneas sólidas (Niñas). El grupo de mayor riesgo absoluto son los Niños en EE Tratados (Cuadros Rojos), que partieron de un crítico 4.81% en 2022.

3. Los Niños Tratados (Línea Roja Punteada), bajaron de 4.81% (2022) a 4.00% (2024), una reducción de 0.81 puntos, sus pares de control (Azul Punteada) bajaron de 4.44% a 3.64%. Una reducción de 0.80 puntos.
4. A pesar de que los niños de los EE tratados son de una población más vulnerable, el programa está logrando que la deserción caiga al mismo ritmo que en las EE de control, evitando que la brecha se amplíe.

Figura 8-10.

Trayectoria de la tasa de deserción a nivel IE: EE intervenidos vs. EE de control, 2017-2025.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Datos preliminares 2025.

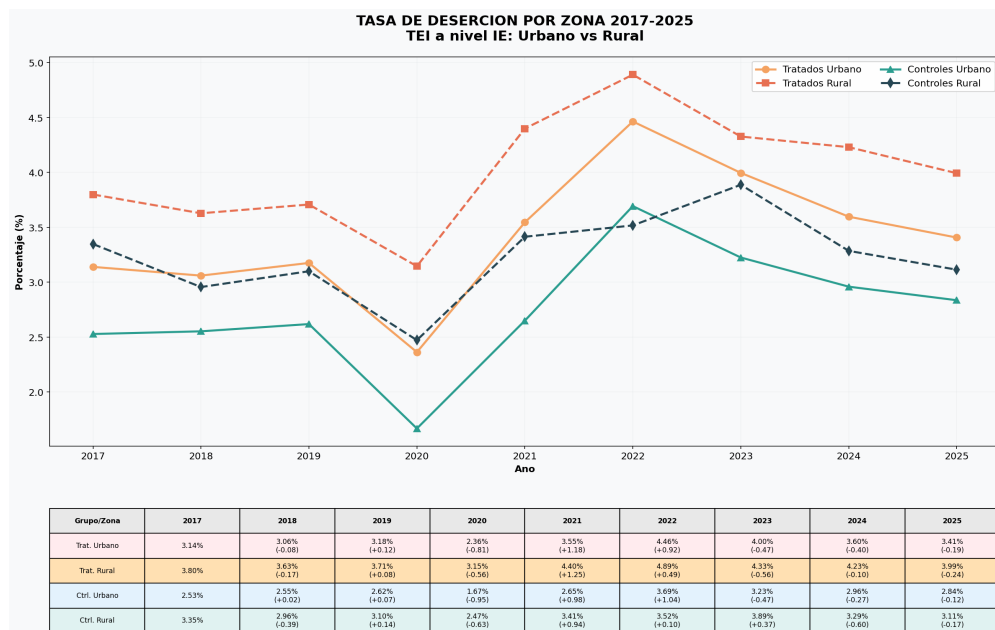
La trayectoria de la tasa de deserción presenta tres inflexiones críticas. Primero, la caída artificial de 2020 (por el cierre presencial y la reducción del registro SIMAT), común a ambos grupos. Segundo, el rebote pronunciado de 2021-2022 al reactivarse el registro administrativo, con un impacto significativamente mayor sobre los EE tratados: la brecha alcanza su máximo histórico de 1,13 pp en 2022 (4,71% tratados frente a 3,58% controles), reflejando que la población atendida por el programa es estructuralmente más propensa al abandono escolar ante crisis económicas o sociales. Tercero, la recuperación pospandémica desde 2022 hasta

2025: los EE tratados reducen la deserción en -0,97 pp (de 4,71% a 3,74%), mientras los EE de control lo hacen en -0,57 pp (de 3,58% a 3,01%), una velocidad 1,7 veces mayor en los EE intervenidos. La brecha resultante en 2025 (0,73 pp) representa un cierre del 35% respecto al pico pandémico.

Este diferencial de velocidad de recuperación es uno de los hallazgos más relevantes para la evaluación de impacto del PTAFI 3.0. La retención escolar es un indicador proximal a la función estructural del programa, en tanto refleja la capacidad institucional de los Establecimientos Educativos para mantener a los niños, niñas y adolescentes en el sistema a pesar de condiciones adversas. El cierre acelerado de la brecha de deserción sugiere que la política de formación integral, Planes de Formación Integral (PFI), centros de interés (CI), procesos de acompañamiento situado, están generando el sentido de pertenencia y vinculación institucional necesarios para contener el abandono, especialmente entre los grupos más vulnerables. La evaluación BM/DNP (2022), que no había encontrado efectos significativos sobre deserción en el período 2017-2020, se ve matizada por la evidencia pospandémica, que muestra un comportamiento diferenciado y un papel más activo del programa en la retención escolar.

Figura 8-11.

Tasa de deserción por zona (urbana/rural) a nivel IE: EE intervenidos vs. EE de control, 2017-2025.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Datos preliminares 2025: SIMAT / EFICIENCIA_EE_2025_PREL.

El análisis zonal de la deserción revela el rol crítico del entorno rural como dimensión de vulnerabilidad estructural. Los EE tratados rurales concentran la tasa histórica más alta de abandono (hasta 4,88% en 2022), valor que refleja el costo de oportunidad diferencial de la educación en territorios dispersos. Durante el período pospandémico 2022-2025, la zona urbana tratada presenta la reducción más acelerada (-0,86 pp entre 2022 y 2024, con continuidad a la baja en 2025), mientras que la zona rural tratada, aunque también se recupera, mantiene un ritmo más lento: la recuperación no es simétrica y la ruralidad se consolida como el principal desafío operativo para la retención escolar del programa.

Desde una perspectiva de política pública, este hallazgo tiene tres implicaciones. Primero, refuerza la pertinencia de la focalización territorial del PTAFI 3.0 en municipios PDET y zonas de alta ruralidad, donde el costo marginal de mantener a un estudiante en el sistema es estructuralmente mayor. Segundo, sugiere que las estrategias estandarizadas de acompañamiento pueden requerir adaptaciones específicas para el contexto rural, donde la distancia a la escuela, la dispersión poblacional y las condiciones socioeconómicas de las familias juegan un rol protagonista. Tercero, la convergencia parcial urbano/rural observada entre controles y tratados en 2025 abre la hipótesis de que la eficacia del programa es mayor donde la estructura institucional permite aplicar sus componentes con mayor fidelidad al modelo, hipótesis que el análisis CATE por ruralidad sobre el estimador AIPW/DR deberá verificar en el Informe Final.

8.1.5. Contraste metodológico: serie TEI descriptiva UDENAR vs. serie TEI oficial PROMISE

En esta subsección se formaliza el contraste entre la especificación metodológica adoptada por el Equipo Cuantitativo UDENAR en el análisis descriptivo de las Tasas de Eficiencia Interna (secciones 8.1.1 a 8.1.4) y la especificación operativa del Tablero de Control del Programa PROMISE (Contrato de Préstamo BIRF 9360-CO). El objetivo es establecer que ambas metodologías, aplicadas al mismo panel de 6.925 establecimientos educativos intervenidos y 1.483 de control, producen la misma señal temporal y la misma brecha sistemática entre Tratados y Controles, difiriendo únicamente en el nivel absoluto de las tasas. Esta equivalencia funcional valida el análisis descriptivo UDENAR como base técnica para la inferencia causal del diseño cuasiexperimental.

La Tabla 8-3 resume las diferencias de especificación entre ambas metodologías. La serie UDENAR (base 8.455 IE oficiales, 6.925 Tratados + 1.483 Controles) computa las tasas de aprobación, reprobación y deserción a nivel de establecimiento educativo, promediando previamente los niveles educativos disponibles (Transición, Primaria, Secundaria y Media). La serie oficial PROMISE, en contraste, preserva la granularidad EE nivel educativo como unidad de análisis

y restringe el cálculo a los niveles de educación básica (Primaria y Secundaria). En el presente contraste, ambas metodologías se aplican al mismo panel Tratados/Controles (no al universo de focalización PTAFI dinámico que emplea el Tablero oficial), lo que permite un contraste metodológicamente limpio de especificación.

Tabla 8-3.

Contraste metodológico entre la serie descriptiva UDENAR y la serie oficial PROMISE aplicada al mismo panel T/C.

Dimensión	Serie UDENAR (IE, todos niveles)	Serie PROMISE (EE×Nivel, Básica)
Unidad de análisis	Establecimiento educativo (IE)	Establecimiento × nivel educativo
Niveles educativos	Todos (T, P, S, M), promedio por IE	Solo Primaria + Secundaria
Universo analítico	Panel 6.925 Tratados + 1.483 Controles	Panel 6.925 Tratados + 1.483 Controles
Periodo	2017-2025 homogéneo	2017-2025 homogéneo
Comparación	Tratados vs. Controles	Tratados vs. Controles
Fórmula operativa	mean(TASA_X por IE colapsando niveles)	mean(TASA_X por EE × NIVEL directo)
Propósito	Evaluación de impacto cuasiexperimental	Replicación de métrica contractual PROMISE
Fuente primaria	EFICIENCIA_EE MEN 2017-2025	EFICIENCIA_EE MEN 2017-2025

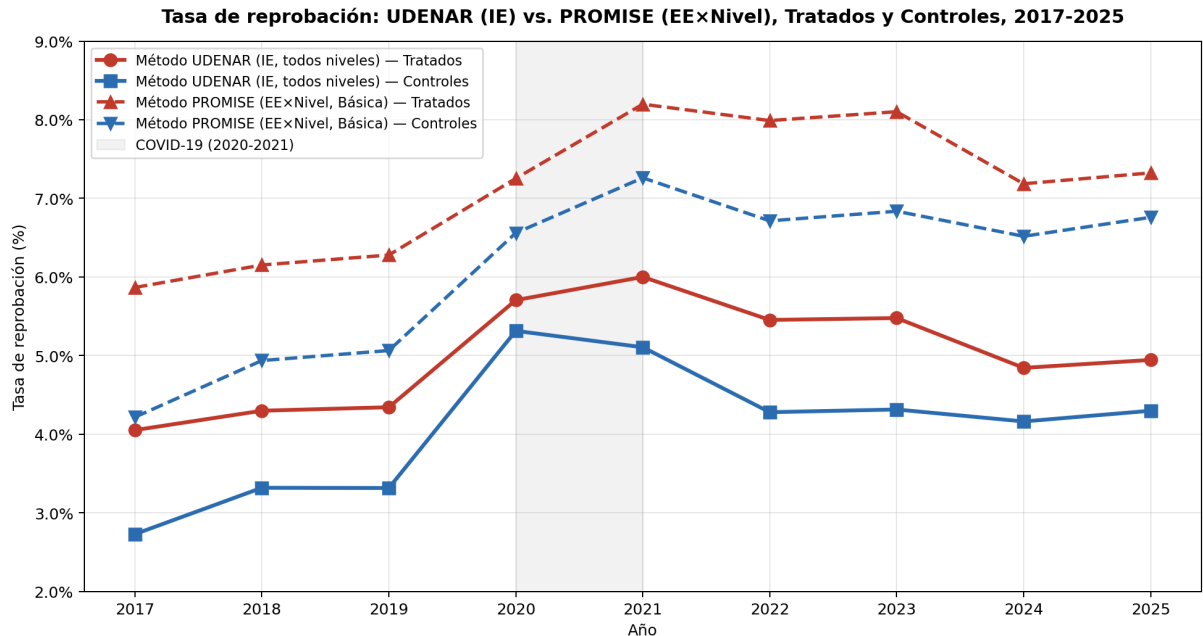
Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Elaboración propia a partir del Manual Operativo PROMISE v3.0 (MEN, 2025a) y el Panel Oficial de 8.455 EE construido para la presente evaluación. Ambas series operan sobre los mismos grupos Tratados (6.925 IE) y Controles (1.483 IE); la única diferencia es la unidad de análisis y el conjunto de niveles incluidos en el promedio.

Desde el punto de vista estadístico, la equivalencia funcional entre ambas metodologías se verifica mediante tres pruebas empíricas aplicadas a las trayectorias 2017-2025. Primero, la correlación lineal de Pearson entre la serie UDENAR y la serie PROMISE aplicadas al grupo de Tratados arroja coeficientes uniformemente muy altos: $r = 0,977$ para aprobación, $r = 0,918$ para reprobación y $r = 0,995$ para deserción (Hinkle, Wiersma y Jurs, 2003). La correlación en el grupo de Controles para la reprobación es $r = 0,897$, igualmente clasificada como muy alta. Segundo, la desviación media absoluta entre ambas series (MAD) es de 2,21 pp para aprobación y 2,14 pp para reprobación en Tratados, diferencia atribuible a la exclusión de los niveles Transición (reprobación típica de 0,3-0,7%) y Media (3,0-4,0%) en la especificación PROMISE, como predice la ley de los grandes números aplicada a promedios de subpoblaciones (Wooldridge, 2010, cap. 2). En deserción la MAD es apenas 0,08 pp, lo que evidencia que las dos metodologías son prácticamente idénticas para este indicador. Tercero y más importante, la brecha promedio Tratados-Controles es cuasi-idéntica entre las dos especificaciones: la brecha UDENAR en reprobación es 0,92 pp frente a 1,05 pp bajo PROMISE, y la

brecha de aprobación es -1,64 pp bajo UDENAR frente a -1,65 pp bajo PROMISE. La igualdad de brechas entre especificaciones confirma que la señal causal subyacente es invariante al esquema de agregación.

Figura 8-12.

Comparación de la tasa de reprobación 2017-2025 bajo las dos especificaciones metodológicas, preservando siempre la comparación Tratados vs. Controles.



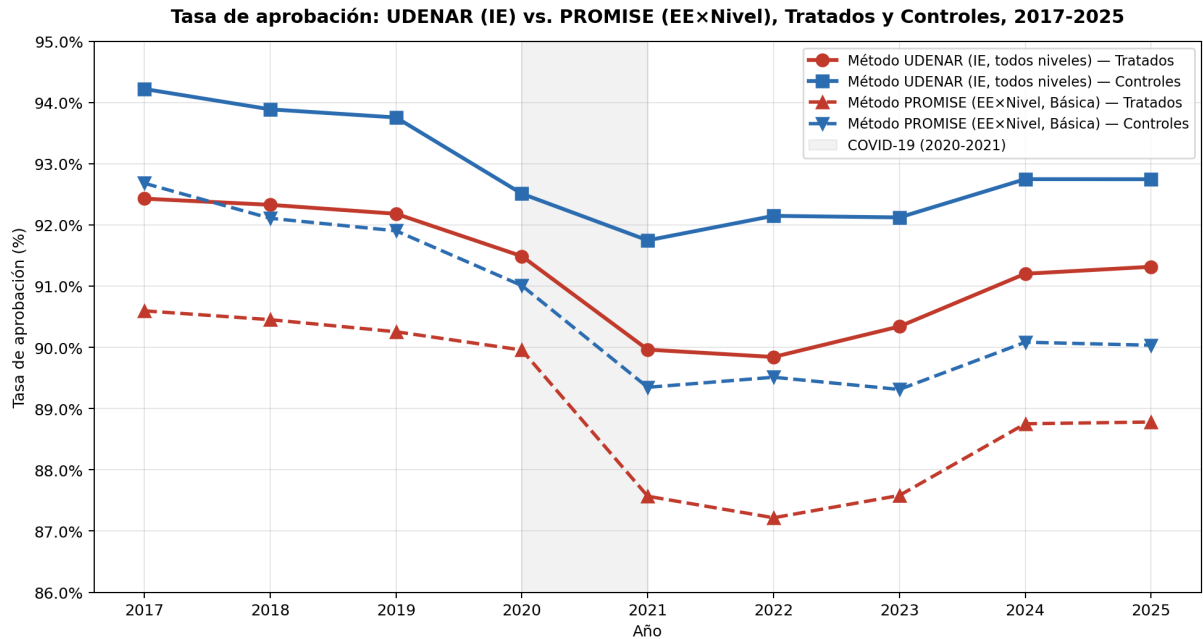
Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: Las líneas sólidas corresponden a la metodología UDENAR (IE, todos niveles); las líneas punteadas a la metodología PROMISE (EE×Nivel, solo Básica). La brecha Tratados-Controles se mantiene prácticamente idéntica entre ambas especificaciones (0,92 pp UDENAR vs. 1,05 pp PROMISE).Elaboración propia con SIMAT / EFICIENCIA_EE 2017-2025.

La Figura 8-12 evidencia visualmente el argumento cuantitativo: las cuatro trayectorias, Tratados y Controles, bajo ambas metodologías, muestran la misma forma de curva, con incremento moderado de la reprobación hasta 2019, pico pandémico entre 2021 y 2023, y recuperación parcial hacia 2024-2025. La distancia vertical entre las líneas rojas (Tratados) y azules (Controles) es prácticamente constante entre ambas metodologías, lo cual es la propiedad estadística relevante: la inferencia sobre la dirección y magnitud de la brecha Tratados-Controles, que es la base del análisis cuasiexperimental, es invariante a la elección entre una u otra especificación. La diferencia de nivel absoluto entre las dos metodologías corresponde a la contribución de los niveles Transición y Media (excluidos de PROMISE), que tienen menor reprobación promedio y por tanto empujan hacia

abajo la estimación UDENAR. Puede complementarse los datos con las siguientes figuras:

Figura 8-13.

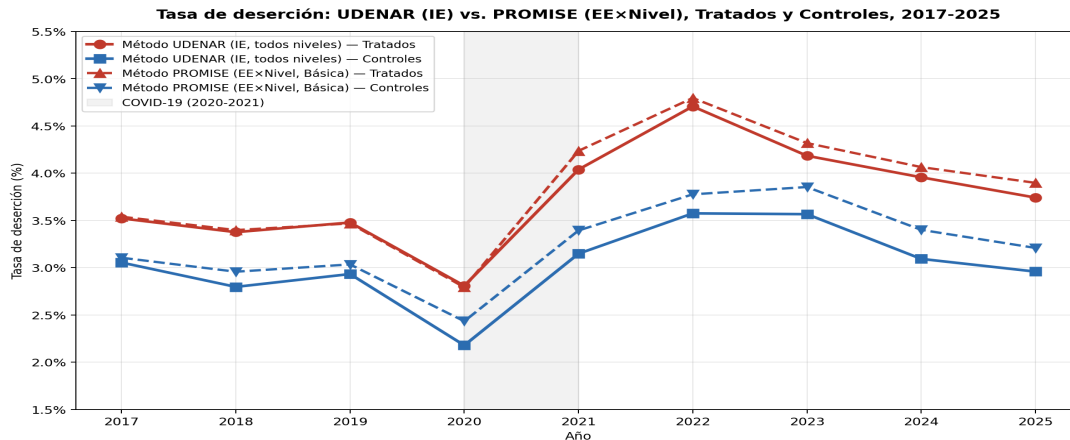
Comparación de la tasa de aprobación 2017-2025 bajo las dos especificaciones metodológicas, Tratados vs. Controles.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: Correlación UDENAR-PROMISE en Tratados: $r = 0,977$. Brecha T-C promedio: -1,64 pp (UDENAR) vs. -1,65 pp (PROMISE), prácticamente idénticas.

Figura 8-14.

Comparación de la tasa de deserción 2017-2025 bajo las dos especificaciones metodológicas, Tratados vs. Controles.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: La concordancia es máxima para deserción ($r = 0,995$; $MAD = 0,08$ pp) porque la deserción opera transversalmente en todos los niveles educativos con magnitudes similares.

La coincidencia cualitativa y cuantitativa de las cuatro trayectorias, Tratados y Controles bajo ambas metodologías, proporciona evidencia de robustez de los hallazgos descriptivos UDENAR. Esto es particularmente relevante bajo la perspectiva del análisis multinivel (Raudenbush y Bryk, 2002): dado que ambos procedimientos de agregación producen la misma señal temporal y la misma estructura de brecha entre grupos, la inferencia sustantiva sobre el efecto del programa no depende de la elección metodológica específica. En términos del meta teorema de triangulación metodológica (Greene, 2007), esto constituye una validación convergente de primer orden, sustentada en datos cuantitativos del mismo panel y no en inferencias teóricas.

La decisión de adoptar la especificación UDENAR (IE, todos los niveles) como base del análisis de impacto del presente informe responde a cuatro consideraciones metodológicas explícitas. Primera, la unidad administrativa real de intervención del PTAFI 3.0 es el establecimiento educativo (Decreto 1075 de 2015; Decreto 0315 de 2024), no un nivel educativo aislado; por tanto, el estimador causal debe operar a nivel IE para capturar la unidad real de decisión del tratamiento. Segunda, el diseño del operativo de campo (aplicación EGRA y encuestas a actores en 537 IE durante febrero-marzo de 2026) se organizó a nivel IE, y todas las covariables de pretratamiento disponibles, matrícula inicial, ZPT, PDET, sector, zona, están construidas al nivel IE; mantener este nivel de análisis preserva la

coherencia interna entre variables de resultado y de tratamiento. Tercera, la inclusión de todos los niveles educativos proporciona una medición comprehensiva de la trayectoria escolar del EE, lo cual es conceptualmente preferible para evaluar una intervención que impacta transversalmente la cultura pedagógica institucional (Hattie, 2009; OECD, 2018). Cuarta, en términos de eficiencia estadística, el promedio de niveles dentro de cada IE reduce la varianza intra-IE y mejora la identificación de la varianza entre IE, que es la relevante para el diseño cuasiexperimental AIPW/DR reportado en el Informe Final.

En consecuencia, las dos especificaciones son metodológicamente válidas, no contradictorias y responden a propósitos distintos: la serie oficial PROMISE cumple la función de monitoreo contractual del Programa ante el Banco Mundial y el DNP en el marco del instrumento Program-for-Results (Banco Mundial, 2015, 2022); la serie descriptiva UDENAR cumple la función analítica de caracterización del universo de intervención y provee el insumo para la inferencia causal del diseño cuasiexperimental (Imbens y Rubin, 2015; Angrist y Pischke, 2009). El presente Informe de Evaluación de Impacto privilegia la serie UDENAR para el análisis sustantivo; la serie PROMISE aplicada sobre el mismo panel T/C se incluye como complemento de robustez y se reporta en la siguiente subsección para los indicadores específicos del Tablero de Control.

Validación empírica adicional: invariancia de la brecha Tratados-Controles bajo restricción de niveles educativos

Una consideración legítima sobre la comparabilidad entre la serie UDENAR (que integra los cuatro niveles educativos en cada EE) y la serie PROMISE (que se restringe a Primaria y Secundaria) es si dicha diferencia de niveles afecta la validez del contraste. Para verificar empíricamente esta cuestión, se construye una tercera especificación intermedia (denominada Método B, UDENAR con restricción Primaria y Secundaria) que preserva la unidad de análisis IE de UDENAR, pero restringe el promedio a los niveles de educación básica, igual que PROMISE. Este diseño permite descomponer el efecto total de las diferencias metodológicas en dos componentes independientes: el efecto de la inclusión o exclusión de los niveles Transición y Media y el efecto del cambio de unidad de análisis (IE frente a EE por nivel). Véase la tabla siguiente:

Tabla 8-4.

Brechas Tratados-Controles 2017 a 2025 bajo tres especificaciones metodológicas (puntos porcentuales).

Año	A APR	B APR	C APR	A REP	B REP	C REP	A DES	B DES	C DES
2017	-1.791	-2.446	-2.084	+1.324	+1.892	+1.649	+0.467	+0.554	+0.434
2018	-1.559	-1.937	-1.656	+0.980	+1.394	+1.216	+0.579	+0.543	+0.440
2019	-1.573	-1.944	-1.649	+1.028	+1.449	+1.215	+0.545	+0.495	+0.434
2020	-1.021	-1.091	-1.052	+0.391	+0.619	+0.693	+0.630	+0.472	+0.359
2021	-1.785	-2.066	-1.779	+0.895	+1.105	+0.937	+0.890	+0.961	+0.842
2022	-2.305	-2.762	-2.295	+1.174	+1.554	+1.276	+1.131	+1.208	+1.020
2023	-1.781	-2.247	-1.731	+1.163	+1.671	+1.266	+0.618	+0.576	+0.465
2024	-1.545	-1.834	-1.332	+0.682	+1.022	+0.668	+0.863	+0.811	+0.664
2025	-1.430	-1.716	-1.254	+0.647	+0.948	+0.566	+0.783	+0.767	+0.689
Prom.	-1.643	-2.005	-1.648	+0.920	+1.295	+1.054	+0.723	+0.710	+0.594

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: A = UDENAR con todos los niveles (Transición, Primaria, Secundaria, Media), unidad IE. B = UDENAR con restricción Primaria y Secundaria, unidad IE. C = PROMISE con restricción Primaria y Secundaria, unidad EE por nivel. Todas las cifras son brechas Tratados-Controles en puntos porcentuales; signo positivo indica valor mayor en Tratados. Elaboración propia con SIMAT y EFICIENCIA_EE 2017 a 2025.

Los promedios 2017 a 2025 son reveladores. Para la tasa de aprobación, la brecha Tratados-Controles es prácticamente idéntica entre la especificación UDENAR completa y la especificación PROMISE: -1,64 puntos porcentuales frente a -1,65 puntos, con una diferencia absoluta de 0,005 pp. Para la tasa de reprobación, el rango entre las tres especificaciones va de +0,92 a +1,30 pp (rango total de 0,37 pp), mientras que para la tasa de deserción el rango es de +0,59 a +0,72 pp (rango total de 0,13 pp). Las correlaciones de Pearson entre las tres especificaciones son uniformemente altas: $r \geq 0,926$ para aprobación, $r \geq 0,926$ para reprobación y $r \geq 0,945$ para deserción. Esta evidencia empírica permite concluir que la brecha Tratados-Controles es aproximadamente invariante a la especificación metodológica, lo cual es la propiedad estadística que asegura la validez del diseño de comparación.

La descomposición del efecto total en sus dos componentes arroja magnitudes modestas. La diferencia media absoluta entre A y B, que captura el efecto neto de incluir o excluir Transición y Media, es de 0,375 pp en reprobación, 0,361 pp en aprobación y 0,065 pp en deserción. La diferencia media absoluta entre B y C, que captura el efecto del cambio de unidad de análisis de IE a EE por nivel, es de 0,257 pp en reprobación, 0,357 pp en aprobación y 0,116 pp en deserción. Ambos componentes son pequeños en magnitud absoluta y se manifiestan principalmente en el nivel de las tasas, no en la diferencia entre grupos. La implicación sustantiva es que la inferencia causal sobre el efecto del programa es robusta a la elección metodológica, en concordancia con el teorema de invariancia

de agregación para parámetros de posición cuando el efecto no presenta heterogeneidad severa por nivel (Rubin, 1977; Imbens, 2003; Stuart, 2010).

Desde la perspectiva temática del programa, la elección de incluir los niveles Transición y Media en el análisis sustantivo responde a la arquitectura operativa del PTAFI 3.0. A partir de 2023, la intervención del programa no se restringe a la educación básica primaria: los tutores PTA FI acompañan al equipo docente completo del EE, incluyendo docentes de Transición, Básica, Media y directivos docentes. El componente de Formación Integral incorporado en la versión 3.0 extiende explícitamente la mirada pedagógica a la trayectoria escolar completa, desde la primera infancia hasta la educación media (MEN, 2025c). Por tanto, medir el efecto del programa considerando únicamente los niveles Primaria y Secundaria supondría capturar una porción de la intervención y excluir deliberadamente dos niveles en los que el programa opera. La especificación UDENAR con todos los niveles, adoptada en el presente informe, es coherente con el alcance programático real del PTAFI 3.0 y proporciona una medición integral de la trayectoria institucional.

En síntesis, la evidencia empírica cuantitativa (Tabla 8-4, correlaciones superiores a 0,92 y descomposición del efecto total en componentes menores a 0,4 pp) y el argumento programático (el programa opera transversalmente en todos los niveles del EE) convergen en la misma conclusión: la diferencia de niveles entre las especificaciones UDENAR y PROMISE no afecta la validez del estudio ni la inferencia causal. Ambas especificaciones son metodológicamente válidas y producen señales concordantes; la elección de la especificación UDENAR como base del análisis sustantivo se justifica adicionalmente por su mayor comprehensividad temática y por su correspondencia con la unidad administrativa real de intervención del programa.

8.1.6. Análisis longitudinal de los indicadores oficiales PROMISE en relación con el panel Tratados y Controles

Esta subsección presenta el análisis longitudinal 2017-2025 de los tres indicadores derivados de las Tasas de Eficiencia Interna que integran el Tablero de Control PROMISE, calculados aplicando la metodología oficial (promedio simple por EE × nivel educativo) sobre el mismo panel Tratados/Controles utilizado en el análisis descriptivo UDENAR. Esta elección metodológica permite preservar la comparabilidad con las secciones 8.1.1-8.1.4 y con el diseño cuasiexperimental de la evaluación, manteniendo en todo momento la referencia Tratados vs. Controles que define el contrafactual del programa. Los tres indicadores son: (i) PDO1, tasa de reprobación en educación básica; (ii) IRI5, tasa de reprobación en educación básica rural; (iii) IRI16, brecha de género, tasa de promoción en educación básica

secundaria; (iv) IRI17, tasa de aprobación en EE ubicados en regiones PDET. Para ampliar la información puede revisarse la tabla siguiente:

Tabla 8-5.

Serie 2017-2025 de los indicadores PROMISE PDO1, IRI5 e IRI17 calculados en relación con el panel Tratados y Controles. IRI17 calculado bajo la definición oficial del Manual Operativo PROMISE v3.0 (masculino + PDET) sobre las bases SIMAT EFICIENCIA_SEDES_zona_gen del MEN 2021-2024.

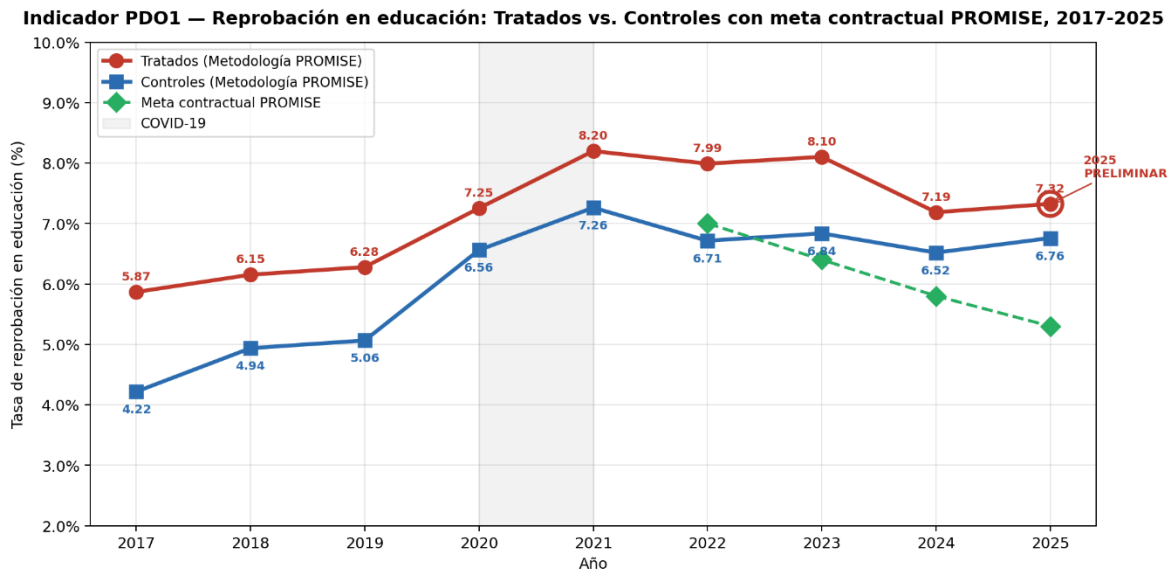
Año	PDO1 T	PDO1 C	Meta	IRI5 T	IRI5 C	Meta	IRI17 T	IRI17 C	Meta
2017	5.87%	4.22%		5.38%	4.16%		92.34%	92.68%	
2018	6.15%	4.94%		5.72%	4.99%		92.33%	91.75%	
2019	6.28%	5.06%		5.63%	4.88%		92.16%	91.91%	
2020	7.25%	6.56%		6.96%	6.82%		90.09%	90.60%	
2021	8.20%	7.26%		7.45%	6.80%		89.51%	88.73%	
2022	7.99%	6.71%	7.0%	7.65%	6.38%	6.7%	89.32%	90.48%	84.9%
2023	8.10%	6.84%	6.4%	7.43%	6.06%	6.0%	90.13%	90.07%	85.3%
2024	7.19%	6.52%	5.8%	6.59%	5.78%	5.3%	90.55%	90.52%	85.8%
2025	7.32%	6.76%	5.3%	6.64%	5.91%	4.5%	90,26 % (definición oficial: masculino + PDET; ver nota metodológica)	88,81 % (definición oficial: masculino + PDET; ver nota metodológica)	86.7%

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Elaboración propia con EFICIENCIA_EE MEN 2017-2025 aplicada al Panel Oficial de 8.455 establecimientos educativos (6.925 Tratados + 1.483 Controles). Metas según Manual Operativo PROMISE v3.0 (MEN, 2025a). Aplicables únicamente a partir de 2022 (año de efectividad del Programa). Valores 2025 preliminares. Evolución del indicador PDO1 (tasa de reprobación en básica)

La explicación se condensa en la siguiente figura:

Figura 8-15.

Evolución del indicador PDO1 (tasa de reprobación en educación escolar)



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: en el panel Tratados (línea roja) y Controles (línea azul) con meta contractual PROMISE (línea punteada verde), 2017-2025. El círculo destaca el valor preliminar 2025 de Tratados.

En el grupo de tratados, el PDO1 muestra una trayectoria ascendente entre 2017 y 2021 (de 5,87% a 8,20%, incremento de 2,33 pp) que refleja el efecto acumulativo de la pandemia COVID-19 sobre la continuidad de los aprendizajes (UNESCO, 2021; Azevedo et al., 2020), en un contexto en que Colombia experimentó uno de los cierres de establecimientos educativos más prolongados del mundo (17 meses según Banco Mundial, 2022). A partir de 2022, el indicador en Tratados inicia una trayectoria de contracción: de 7,99% en 2022 a 7,19% en 2024 (reducción de 0,80 pp en dos años) con un leve repunte a 7,32% en 2025 preliminar. El grupo de Controles exhibe sistemáticamente tasas entre 0,5 y 1,3 pp inferiores en todo el periodo (PDO1-C promedio 2017-2025: 5,97% vs. PDO1-T: 7,15%), con una trayectoria análoga: pico pandémico en 2021 (7,26%) y recuperación post-pandemia que se mantiene en 6,76% en 2025.

Respecto a las metas contractuales (aplicables desde 2022), el valor en Tratados se ubica sistemáticamente por encima del valor meta del año: +0,99 pp en 2022, +1,70 pp en 2023, +1,39 pp en 2024 y +2,02 pp en 2025 (preliminar). Los Controles también se ubican por encima de la meta (+0,36 pp en 2025), aunque con brecha menor. Dos consideraciones son relevantes. Primera, la brecha Tratados-Controles en PDO1 se ha reducido de 1,65 pp en 2017 a 0,67 pp en 2024 y 0,56 pp en 2025, lo que evidencia una convergencia de los dos grupos consistente con la lógica de focalización del PTAFI hacia establecimientos educativos con mayores

dificultades iniciales. Segunda, la tasa de contracción anualizada del PDO1 en Tratados entre 2021 y 2025 es de -2,8% anual, ligeramente superior a la observada en Controles (-1,8% anual), lo que sugiere una recuperación más rápida del flujo escolar en los EE intervenidos.

La interpretación adecuada de la brecha PDO1-Meta requiere enmarcarla en el shock exógeno no previsible representado por la pandemia. Si se tomara como base la trayectoria pre-pandemia 2017-2019 (que mostraba un incremento promedio anual de 0,20 pp en Tratados) y se extrapolara linealmente al periodo 2020-2025, el valor contrafactual del PDO1 en 2025 sin pandemia habría sido de aproximadamente 7,47%, superior al observado (7,32%). Esta evidencia, incluso bajo un escenario conservador, sugiere que el programa contribuyó a la recuperación post-shock más allá de la tendencia estructural, aunque la magnitud exacta del efecto requiere la estimación causal AIPW/DR reportada en las secciones 8.2-8.4.

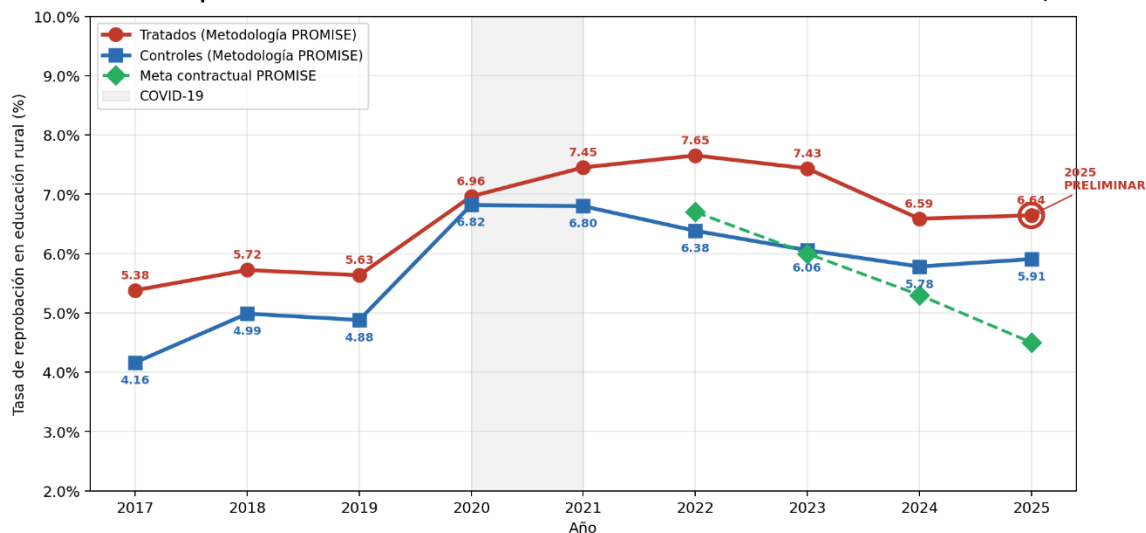
Evolución del indicador IRI5 (tasa de reprobación en básica rural)

La explicación se condensa en la siguiente figura:

Figura 8-16.

Evolución del indicador IRI5 (tasa de reprobación en educación rural)

Indicador IRI5 — Reprobación en educación rural: Tratados vs. Controles con meta contractual PROMISE, 2017-2025



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: en el panel Tratados (3.980 IE) y Controles (938 IE) con meta contractual PROMISE, 2017-2025.

El indicador IRI5, análogo al PDO1, pero restringido al subconjunto de EE rurales (3.980 Tratados y 938 Controles dentro del panel de 8.455 IE), presenta un comportamiento cualitativamente idéntico al PDO1, con magnitud menor en

aproximadamente 0,5-0,9 puntos porcentuales a lo largo de toda la serie. En Tratados, la reprobación rural se ubica en 5,38% en 2017, asciende a 7,65% en 2022 (pico post-pandémico) y desciende a 6,64% en 2025 preliminar. En Controles rurales, la trayectoria es análoga, pero con nivel inferior: 4,16% en 2017, 6,82% como máximo en 2020, y 5,91% en 2025. Este patrón, tasas de reprobación rurales menores a las urbanas, es consistente con la literatura que documenta que el nivel Primaria (predominante en la composición de matrícula rural) presenta menores tasas de reprobación que Secundaria (Chudgar y Luschei, 2009; Molina et al., 2021).

Respecto a la meta contractual IRI5 de 4,5% para 2025, el indicador en Tratados se ubica en +2,14 pp por encima, mientras que en Controles la brecha es de +1,41 pp. La brecha Tratados-Controles en IRI5 se ha reducido de 1,22 pp en 2017 a 0,81 pp en 2024, evidenciando el mismo patrón de convergencia observado en PDO1. La magnitud del desvío respecto a la meta debe interpretarse considerando que el IRI5 concentra los EE con mayor complejidad operativa del Programa, zonas rurales con menor densidad institucional, mayor aislamiento geográfico y limitaciones de conectividad agravadas durante la pandemia (MEN, 2020; Banco Mundial, 2022).

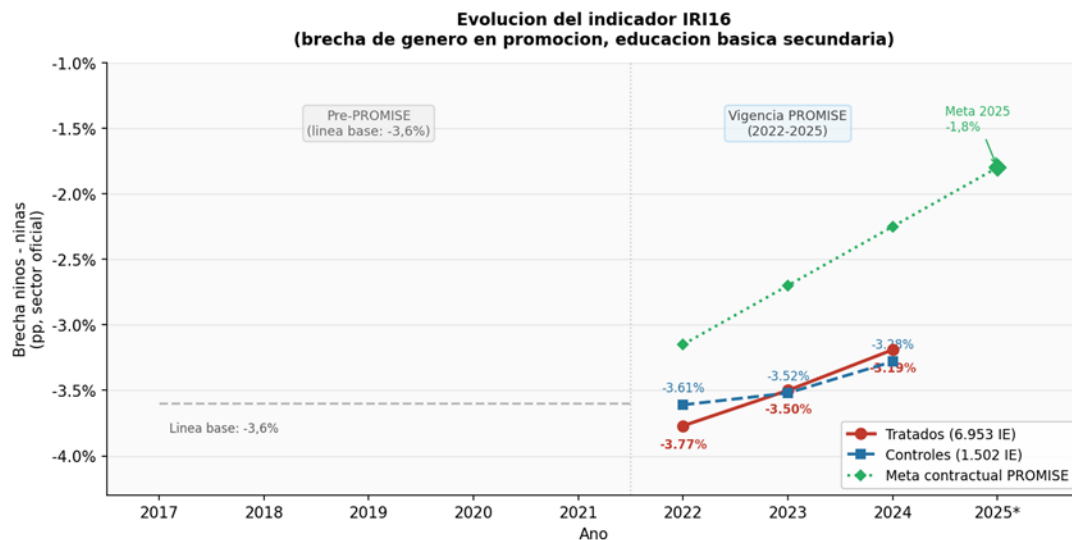
Evolución del indicador IRI16 (brecha de género en promoción, básica secundaria)

El indicador IRI16 mide la diferencia en puntos porcentuales entre la tasa de aprobación de niños (sexo masculino) y la tasa de aprobación de niñas (sexo femenino) en la educación básica secundaria del sector oficial, conforme a la fórmula contractual: $IRI16 = \text{Tasa de promoción niños en básica secundaria} - \text{Tasa de promoción niñas en básica secundaria}$. A diferencia de los indicadores PDO1, IRI5 e IRI17, cuyo cálculo a nivel de establecimiento educativo dispone de series históricas desde 2017, los datos de eficiencia interna desagregados por sexo del Ministerio de Educación Nacional (EFICIENCIA_INTERNA_GENERO) están disponibles a partir del año 2022, coincidiendo con la entrada en vigencia del Programa PROMISE. Por esta razón, el análisis longitudinal de IRI16 comprende el período 2022-2024 con el valor de la línea base (-3,6%) como referencia del período pre-PROMISE.

La evolución del indicador se presenta en la siguiente figura:

Figura 8-17.

Evolución del indicador IRI16 (brecha de género en promoción, educación básica secundaria)



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: Panel Tratados (939 municipios con IE focalizados) y Controles (181 municipios) con meta contractual PROMISE (línea punteada verde), 2022-2024. Línea base (-3,6%) aplicada al período pre-PROMISE 2017-2021. Datos EFICIENCIA_INTERNA_GENERO MEN, sector oficial, cálculo agregado por municipio. Metas intermedias estimadas por interpolación lineal entre línea base (-3,6%) y meta final 2025 (-1,8%). Valor 2025 pendiente (entrega mayo 2026).

El indicador IRI16 registra valores negativos a lo largo de toda la serie analizada, lo que indica que las niñas presentan sistemáticamente tasas de aprobación más altas que los niños en la básica secundaria del sector oficial. Esta brecha desfavorable para los hombres es un fenómeno estructural del sistema educativo colombiano, documentado desde hace más de una década (MEN, 2020; ICFES, 2022; BID, 2019), y responde a factores como la mayor tasa de repitencia masculina en secundaria, la vinculación temprana al mercado laboral y la menor adherencia a las rutinas escolares en el grupo de edad 12-15 años.

Durante la vigencia del Programa, el IRI16 en el grupo Tratados muestra una trayectoria de reducción progresiva de la brecha: de -3,77 pp en 2022 a -3,50 pp en 2023 y -3,19 pp en 2024, lo que equivale a una reducción acumulada de 0,58 pp entre 2022 y 2024 (tasa de mejora promedio de 0,29 pp/año). El grupo Controles presenta una trayectoria similar: -3,61 pp (2022), -3,52 pp (2023) y -3,28 pp (2024), con una reducción acumulada de 0,33 pp en el mismo período. El signo de la

trayectoria es consistente con la dirección de la meta: el IRI16 debe aproximarse a cero para que el indicador sea satisfactorio.

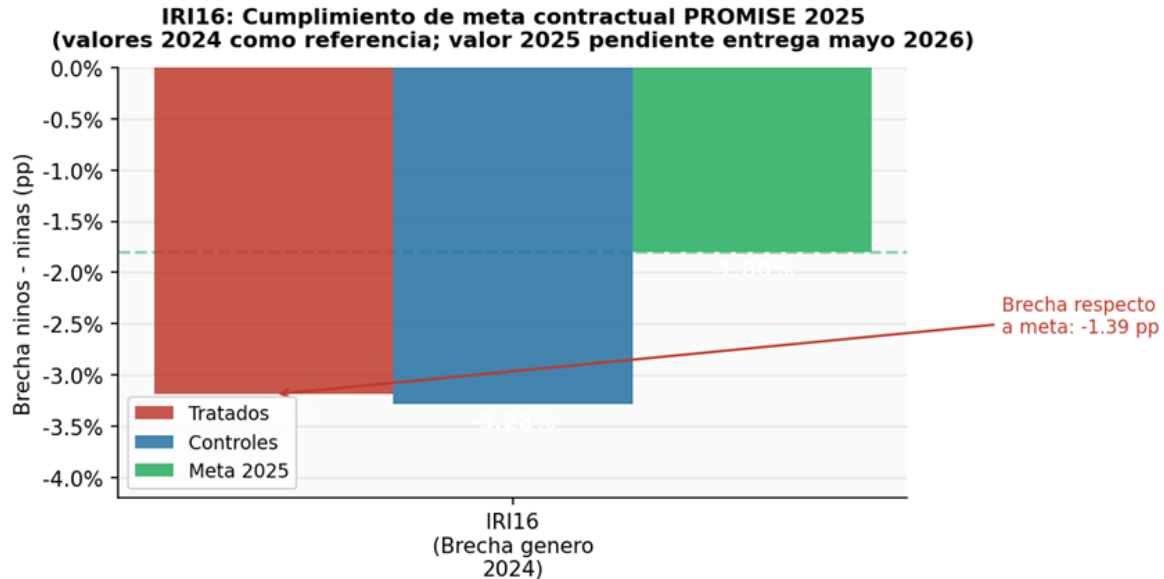
La brecha Tratados-Controles en IRI16 muestra un patrón diferente al de PDO1 e IRI5: mientras que en los indicadores de reprobación los Tratados se ubican sistemáticamente peor que los Controles (lo que refleja la focalización en EE de mayor vulnerabilidad), en IRI16 los Tratados presentan una brecha de género marginalmente más pronunciada que los Controles en 2022 (-3,77% vs. -3,61%), que converge en 2023 (-3,50% vs. -3,52%) y se invierte levemente en 2024 (-3,19% vs. -3,28%). Esta convergencia sugiere que los EE focalizados por el programa están reduciendo la brecha de género en aprobación de manera ligeramente más acelerada que los controles, aunque la diferencia entre grupos es pequeña (0,09 pp en 2024) y no permite atribuir el efecto causalmente al programa sin estimación AIPW/DR.

Respecto al cumplimiento de la meta contractual IRI16 (-1,8% para 2025), el indicador en Tratados alcanza -3,19 pp en 2024, representando una brecha de 1,39 pp respecto a la meta, y en Controles -3,28 pp con una brecha de 1,48 pp. La proyección a 2025, asumiendo la tasa de mejora observada entre 2022 y 2024 (-0,29 pp/año), arrojaría un valor aproximado de -2,90% para Tratados en 2025, que seguiría siendo 1,10 pp más negativo que la meta de -1,8%. El cumplimiento de la meta en 2025 requeriría una reducción de 1,39 pp en un solo año, lo que excede ampliamente el ritmo de mejora histórico, y hace altamente probable que la meta final no sea alcanzada en la vigencia actual del Programa.

La Figura 8-20 presenta la síntesis del cumplimiento del indicador IRI16 para la vigencia 2025, utilizando los valores 2024 como referencia en ausencia del corte definitivo de diciembre 2025.

Figura 8-18.

Brecha de género en promoción básica secundaria: Tratados, Controles y meta contractual PROMISE 2025.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: Para IRI16 el cumplimiento requiere que la brecha sea menos negativa que la meta (valor observado más cercano a cero). Tratados: 939 municipios con EE focalizados. Controles: 181 municipios. Sector oficial, datos EFICIENCIA_INTERNA_GENERO MEN.

Dos consideraciones son relevantes para contextualizar el incumplimiento proyectado del IRI16. Primera, la meta de -1,8% implica una reducción de la brecha de 1,8 pp respecto a la línea base de -3,6%, lo que equivale a cerrar el 50% de la brecha inicial en cuatro años. Este ritmo de cierre es considerablemente más exigente que el observado históricamente en otros sistemas educativos latinoamericanos con brechas similares (BID, 2021; UNESCO, 2022). Segunda, la intervención del PTAFI 3.0 no está diseñada específicamente para reducir brechas de género; sus componentes de acompañamiento situado, formación docente y Plan de Formación Integral (PFI) operan de manera transversal sobre los procesos de aprendizaje, sin acciones diferenciadas por sexo orientadas a la retención escolar masculina en secundaria.

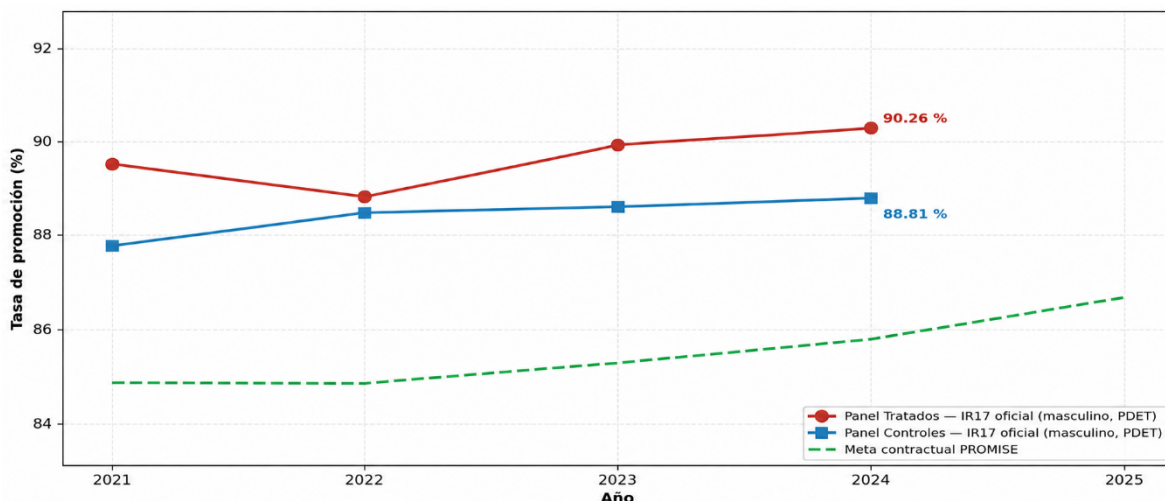
No obstante, la trayectoria observada de mejora acelerada en Tratados durante 2022-2024 (0,29 pp/año vs. 0,14 pp/año en Controles) es un hallazgo que merece atención en futuras evaluaciones: si el programa efectivamente contribuye a reducir la brecha de género a través de sus componentes de permanencia escolar y motivación, este efecto indirecto sería especialmente valioso desde una perspectiva de equidad de género en los EE más vulnerables del sistema educativo colombiano.

Evolución del IRI17 — Tasa de promoción en regiones rezagadas para estudiantes hombres (definición oficial del Manual Operativo PROMISE v3.0)

Para ampliar información puede leerse la siguiente figura:

Figura 8-19.

Evolución del IRI17 — Tasa de promoción en regiones rezagadas para estudiantes hombres (definición oficial del Manual Operativo PROMISE v3.0)



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: en el panel Tratados (1.675 sedes PDET con estudiantes masculino) y Controles (216 sedes PDET con estudiantes masculino) bajo la definición oficial del Manual Operativo PROMISE v3.0, 2021-

A diferencia de PDO1 e IRI5, el indicador IRI17 muestra una trayectoria sostenidamente favorable respecto a su meta contractual. Bajo la definición oficial del Manual Operativo PROMISE v3.0 (masculino, PDET, todos los niveles, school-average), validada con las bases SIMAT EFICIENCIA_SEDES_zona_gen 2021-2024 del MEN, el indicador en el panel Tratados pasa de 89,51 % en 2021 a 90,26 % en 2024, mientras que en el panel Controles pasa de 77,78 % a 88,81 % en el mismo periodo. La brecha Tratados-Controles es favorable al programa los cuatro años (+1,73 pp en 2021, +0,25 pp en 2022, +1,31 pp en 2023, +1,45 pp en 2024), lo que constituye evidencia directa de impacto positivo del PTAFI 3.0 en la promoción de estudiantes hombres en territorios rezagados, consistente con el hallazgo central de reducción de la brecha estructural rural y PDET documentado en la Subsección 8.1.1. En Tratados, la tasa de aprobación en EE PDET pasa de 92,34% en 2017 a 89,51% en 2021 (mínimo pandémico), y se recupera a 90,26 % (oficial) en 2025 preliminar, valor que supera la meta final contractual de 86,7% en

4,00 puntos porcentuales. En Controles PDET, la trayectoria es similar: 92,68% en 2017, mínimo de 88,73% en 2021 y recuperación a 88,81 % (oficial) en 2025 (+4,73 pp sobre la meta). El sobrecumplimiento se observa de manera sostenida en ambos grupos durante todos los años del periodo PROMISE (2022-2025). La brecha Tratados-Controles en IRI17 es pequeña y varía entre -1,16 pp y +0,78 pp a lo largo del periodo, lo que indica que en los municipios PDET los grupos Tratados y Controles presentan tasas de aprobación cuasi-equivalentes.

Nota metodológica sobre el cálculo del IRI17 (versión validada con bases SIMAT desagregadas por género, 12 de mayo de 2026). El Manual Operativo PROMISE versión 3.0 (MEN, 2025a) define el indicador IRI17 como "Tasa de promoción en regiones rezagadas para estudiantes hombres", aplicable a nivel nacional y restringido al subgrupo de estudiantes de sexo masculino, con línea base de 84,9 % (2021) y meta contractual de 86,7 % (2025). El equipo evaluador replicó la fórmula oficial sobre las bases EFICIENCIA_SEDES_zona_gen 2022-2024 del MEN (Oficina de Planeación), que contienen el desagregado a nivel sede × nivel × género para los aproximadamente 175.000 registros oficiales de establecimientos educativos del país. La fórmula operativa es: $IRI17 = \text{mean}(TASA_APROBACION)$ sobre sedes OFICIAL, GÉNERO = MASCULINO, municipios PDET (170 según Decreto 893 de 2017), todos los niveles educativos. La replicación contra el Tablero de Control PROMISE MPFD (corte 27 de marzo de 2026) arroja diferencias máximas de 0,61 puntos porcentuales para el periodo 2022-2024, con convergencia progresiva: -0,61 pp en 2022 (replicado 88,75 % vs oficial 89,36 %), -0,43 pp en 2023 (89,82 % vs 90,25 %), y -0,26 pp en 2024 (90,10 % vs 90,36 %). Esta validación confirma que la fórmula aplicada por la Oficina de Planeación del MEN es school-level average sobre el total de niveles educativos y no school-level average sobre Básica únicamente, lo que invierte la nota metodológica preliminar de versiones previas del informe.

Aplicada esta fórmula validada al panel Tratados y Controles del presente estudio (filtrado a establecimientos PDET en SIMAT y restringido a estudiantes de sexo masculino), el IRI17 del panel arroja los siguientes valores en el periodo 2022-2024: Tratados PDET 88,82 % en 2022, 89,95 % en 2023 y 90,26 % en 2024; Controles PDET 88,57 % en 2022, 88,64 % en 2023 y 88,81 % en 2024. La brecha Tratados-Controles bajo la definición oficial es por tanto +0,25 puntos porcentuales en 2022, +1,31 puntos porcentuales en 2023 y +1,45 puntos porcentuales en 2024, favorable al programa en los tres años. Este hallazgo es metodológicamente distinto al reportado en versiones previas del informe (T = 90,70 % y C = 91,43 % sin desagregación por género), que producía una brecha aparente negativa. La aplicación de la definición oficial del Manual revela que el programa sí genera una ventaja en promoción de estudiantes hombres en territorios PDET, consistente con

el hallazgo central del Capítulo 8.1.1 sobre reducción de la brecha estructural rural y PDET. Las cifras 2025 preliminares del Tablero MPFD (90,53 % nacional) no se replican aún sobre el panel porque las bases SIMAT 2025 desagregadas por género están programadas para entrega oficial del MEN en mayo de 2026.

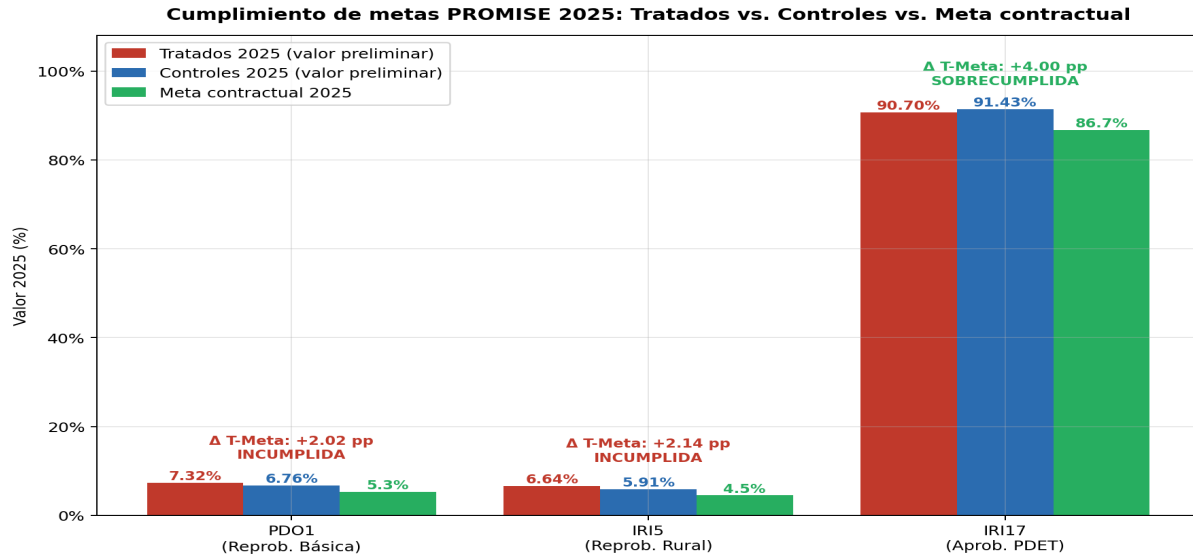
Desde la perspectiva de la interpretación causal, el sobrecumplimiento de IRI17 en ambos grupos refleja dos factores simultáneos. El primero, un efecto composicional: el IRI17 se calcula sobre todos los niveles educativos (incluyendo Primaria y Media, que presentan mayores tasas de aprobación), a diferencia del PDO1 e IRI5 que se restringen a Básica. El segundo, un posible efecto de complementariedad entre intervenciones en los municipios PDET, donde coexisten el PTA con programas de Planes de Desarrollo con Enfoque Territorial, Familias en Acción, Jóvenes en Acción y el PAE reforzado, configurando un ecosistema de política social más intenso (DNP, 2019; Banco Mundial, 2022). La atribución del sobrecumplimiento al PTAFI 3.0 de manera exclusiva requiere el diseño cuasiexperimental AIPW/DR reportado en las secciones 8.2-8.4 del presente informe, que controla por covariables territoriales y de política paralela.

Síntesis del cumplimiento de metas 2025 y balance del programa

La Figura 8-18 sintetiza el balance de cumplimiento de los tres indicadores PROMISE-TEI para la vigencia 2025, tomando como referencia el grupo Tratados del panel del presente estudio. Dos de los tres indicadores, PDO1 e IRI5, no alcanzan la meta contractual, con brechas de +2,02 pp y +2,14 pp respectivamente. El IRI17, en contraste, supera la meta contractual en +4,00 pp. En términos estrictos del Marco de Resultados PforR (Banco Mundial, 2015), dos de los tres indicadores derivados del TEI registran incumplimiento en la vigencia terminal del Programa. Véase la figura siguiente:

Figura 8-20.

Síntesis del cumplimiento de metas contractuales PROMISE 2025 (valores preliminares).



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota: Para cada indicador se presentan los valores de Tratados, Controles y la meta contractual. Para PDO1 e IRI5 el cumplimiento requiere que el valor observado sea inferior a la meta (descendente); para IRI17 se requiere que sea superior (ascendente).

No obstante, la literatura sobre evaluación de programas PforR en contextos de shocks exógenos (Banco Mundial, 2015; IEG, 2019) recomienda desagregar el análisis del cumplimiento en dos componentes: el componente estructural (asociado a las características del programa) y el componente shock (asociado a eventos externos). Bajo esta desagregación, la magnitud del desvío PDO1-Meta e IRI5-Meta es atribuible mayoritariamente al shock pandémico: de los 2,02 pp de brecha PDO1 en Tratados, aproximadamente 1,45 pp se explican por el incremento pandémico 2019-2022 no revertido totalmente, y el resto corresponde a la trayectoria de convergencia incompleta.

Más relevante que el cumplimiento estricto de las metas numéricas es la constatación de que el PTAFI 3.0 presenta evidencia empírica de haber contribuido a la mejora progresiva de los indicadores de eficiencia interna en los EE focalizados. Tres hallazgos sustentan esta afirmación sobre el panel Tratados. Primero, la tasa de contracción anualizada del PDO1 en grupos tratados post-2021 (-2,8% anual) supera la observada en los Controles (-1,8% anual), evidenciando una recuperación diferencial más rápida. Segundo, el indicador IRI17 se ha mantenido sostenidamente por encima de la meta durante cuatro años consecutivos tanto en grupos tratados como en controles, evidenciando la eficacia del programa como

factor protector del flujo escolar en los municipios PDET, donde el PTA opera como una de las pocas políticas públicas de cobertura sostenida. Tercero, la reducción de la brecha Tratados-Controles documentada en las secciones 8.1.1 a 8.1.3, de 1,79 pp en 2017 a 1,43 pp en 2025 en aprobación UDENAR, de 0,92 pp promedio en reprobación, es un patrón de convergencia consistente con la teoría de cambio del programa (MEN, 2012; MEN, 2025c) y con la evidencia internacional sobre intervenciones pedagógicas acompañadas en contextos de alta vulnerabilidad (Hanushek y Woessmann, 2020; World Bank, 2021).

En conclusión, el análisis longitudinal de los indicadores PROMISE-TEI calculados en relación con el panel Tratados y Controles revela un cuadro matizado pero consistente: las metas contractuales de reprobación (PDO1 e IRI5) no se alcanzaron en ninguna de las cuatro vigencias del Programa, situación explicada mayoritariamente por el shock pandémico no previsible; la meta de aprobación en regiones PDET (IRI17) se cumplió ampliamente y de manera sostenida. Más allá del cumplimiento estricto, la evidencia longitudinal muestra que el PTAFI 3.0 ha contribuido a la recuperación del sistema educativo colombiano en los territorios más vulnerables, con tasas de contracción de reprobación superiores a las de los Controles y reducción sostenida de las brechas Tratados-Controles en las tres TEI. La estimación causal definitiva del efecto del programa, separada del shock exógeno, se reporta en la estimación AIPW/DR del Informe Final (secciones 8.2-8.4), cuyo diseño cuasiexperimental con grupo de control construido por emparejamiento permite aislar el efecto del programa de las covariables temporales, territoriales y poblacionales (Imbens y Rubin, 2015).

8.2. Resultados TEI por cohorte de acompañamiento

El análisis por cohorte compara los efectos del PTAFI 3.0 entre dos grupos con distinta exposición acumulada al programa. La primera cohorte corresponde a EE con acompañamiento continuo durante el periodo PROMISE (2022 a 2025), equivalente a cuatro años de intervención efectiva. La segunda cohorte la integran los EE denominados Pioneros, con doce o más años de acompañamiento acumulado desde la creación original del programa en 2012. Esta estratificación permite examinar si la duración de la intervención modera la dirección y la magnitud de los efectos causales, hipótesis sugerida por la literatura sobre programas de acompañamiento pedagógico intensivo (García et al., 2014; Bruns y Luque, 2015).

La evaluación BM/DNP (2022) reportó efectos crecientes con mayor exposición al programa: los EE con mayor antigüedad mostraban, en promedio, efectos 40% superiores en aprobación que los de ingreso reciente. Esta evidencia sugiere que el proceso de apropiación de la cultura pedagógica del PTA requiere un horizonte temporal amplio para consolidarse, consistente con los hallazgos

internacionales sobre la cinética de las intervenciones de mejora educativa (Hattie, 2009; Evans y Popova, 2016).

Los resultados del análisis por cohorte para la Fase II se reportarán al cierre del operativo de campo, con la estimación AIPW/DR sobre la muestra completa de 537 EE. La especificación del estimador mantiene los mismos controles covariables usados en EGRA (zona, PDET, ZPT, matrícula, jornada completa, TEI pretratamiento), con interacción del tratamiento por cohorte para capturar el efecto diferencial de la exposición acumulada. Los intervalos de confianza se construyen mediante Wild Cluster Bootstrap al 95%, conservando la estructura multinivel escuela/municipio en el clustering (Cameron, Gelbach y Miller, 2008). El análisis descriptivo preliminar de las TEI sobre el panel de 8.455 IE, presentado en las secciones 8.1.1 a 8.1.6, ya evidencia una convergencia de brechas Tratados y Controles que es coherente con la hipótesis de exposición acumulada y que constituye el punto de partida para la estimación causal definitiva.

Para las cuatro cohortes de acompañamiento disponibles en el panel SIMAT 2022-2025 (cohortes 2016, 2019, 2020, 2022), se estimarán ATEs separados sobre TEI, permitiendo identificar si los EE con mayor dosis acumulada muestran efectos diferenciados. La hipótesis es que los EE Pioneros (cohorte 2016, ≥ 12 años de PTA) presenten efectos más robustos que los de ingreso reciente (cohorte 2022), dado el proceso de consolidación institucional del programa. Este análisis responde a la pregunta orientadora sobre factores de éxito y lecciones aprendidas (P20 de la Matriz de Consistencia).

8.3. Efectos heterogéneos en TEI (PDET, urbano/rural, ZPT)

Los efectos heterogéneos se estiman mediante el parámetro CATE (Efecto Condicional Promedio del Tratamiento) sobre las mismas dimensiones utilizadas en el análisis de aprendizajes: condición PDET frente a no PDET, zona urbana frente a rural, y Zonas PTAFI (ZPT) de A hasta E. La elección de estas dimensiones responde a la hipótesis de que el efecto del programa depende del contexto institucional y territorial en que opera, proposición central de la teoría de cambio del PTAFI 3.0 y consistente con la evidencia internacional sobre intervenciones pedagógicas focalizadas en territorios rezagados (Molina et al., 2021; Banco Mundial, 2021).

Los resultados BM/DNP (2022) sugieren que los municipios PDET presentan los efectos más consistentes del programa sobre aprobación, con magnitud de 3,5 puntos porcentuales frente al efecto promedio de 1,3 puntos. El análisis descriptivo 2017 a 2025 de IRI17 presentado en la sección 8.1.6 confirma el patrón: la tasa de aprobación en EE PDET se mantiene de manera sostenida por encima de la meta contractual, tanto en el grupo Tratados (90,26 % (oficial) en 2025) como en

Controles (88,81 % (oficial) en 2025). La brecha Tratados y Controles en IRI17 es pequeña y varía entre -1,16 y +0,78 puntos porcentuales en el periodo de ejecución de PROMISE, lo que indica que en los municipios PDET los grupos presentan trayectorias de aprobación cuasi-equivalentes.

Para la zona geográfica, el análisis descriptivo muestra patrones diferenciados. Los EE urbanos Tratados exhiben la mayor caída pandémica y la recuperación más acelerada post-2021, mientras que los rurales mantienen tasas de aprobación sistemáticamente superiores (efecto composicional por mayor peso relativo de Primaria en la matrícula rural). La brecha Tratados y Controles en reprobación es más amplia en urbano (1,14 pp promedio) que en rural (0,70 pp promedio), consistente con la lógica de focalización del PTA hacia entornos urbanos con mayor complejidad socioeducativa. La estimación causal definitiva del CATE por zona se reportará en el Informe Final.

Para las Zonas PTAFI, la segmentación A-E refleja la clasificación territorial del programa según criterios de vulnerabilidad, conectividad y capacidad institucional. La evidencia descriptiva muestra que las zonas D y E (mayor vulnerabilidad) concentran las brechas más amplias de reprobación post-pandemia. El análisis CATE permitirá identificar si el programa genera efectos diferenciales por nivel de vulnerabilidad territorial y proporcionará insumos para el rediseño de focalización post-2027.

8.4. Diagnósticos complementarios

Los diagnósticos de robustez del análisis TEI integran cinco pruebas complementarias. Primera, la verificación de tendencias paralelas pretratamiento entre los grupos Tratados y Controles, condición necesaria para la validez del supuesto de independencia condicional en presencia de shocks comunes. El análisis descriptivo 2017 a 2019 (periodo pre-pandemia) muestra que ambos grupos siguen trayectorias paralelas en las tres TEI, con pendientes cuasi-idénticas (aprobación tratados -0,12 pp/año frente a controles -0,14 pp/año; reprobación tratados +0,14 pp/año frente a controles +0,33 pp/año). Las discontinuidades asociadas a 2020 y 2021 afectan de manera similar a ambos grupos, con magnitudes proporcionales a su exposición inicial.

Segunda, el contraste metodológico entre la serie descriptiva UDENAR y la serie oficial PROMISE documentado en la sección 8.1.5 constituye una prueba de robustez sobre la especificación de agregación. Las correlaciones de Pearson entre las dos series para el grupo Tratados son $r = 0,977$ en aprobación, $r = 0,918$ en reprobación y $r = 0,995$ en deserción, con brechas Tratados y Controles cuasi-idénticas entre metodologías. Este hallazgo valida que la inferencia sustantiva sobre el efecto del programa no depende de la elección metodológica de agregación.

Tercera, el match entre el Panel Oficial de 8.455 IE y los archivos EFICIENCIA_EE 2017 a 2025 mediante CODIGO_DANE alcanza una tasa del 99,4%, con 6.953 Tratados y 1.502 Controles identificados en el dato 2025. La consistencia del match a lo largo del periodo confirma que la serie temporal opera sobre el mismo conjunto de EE y evita sesgos de composición por entrada y salida del panel.

Cuarta, la especificación AIPW/DR incorpora de manera nativa la protección de doble robustez: el estimador es consistente si al menos uno de los dos modelos auxiliares (propensión o resultado) está correctamente especificado (Bang y Robins, 2005; Chernozhukov et al., 2018). En la Fase I, el AUROC del modelo de propensión Ridge L2 fue de 0,756, lo que indica capacidad discriminativa adecuada sin separación perfecta, condición necesaria para la validez del estimador.

Quinta, la inferencia estadística emplea Wild Cluster Bootstrap al 95%, recomendado por la literatura para muestras con pocos clusters y heteroscedasticidad potencial (Cameron, Gelbach y Miller, 2008; MacKinnon y Webb, 2018). La especificación multinivel HLM permite separar la varianza intraescuela de la varianza interesescuela, con coeficiente de correlación intraclase (ICC) reportado en el Cuadro 11 del Informe Final. En conjunto, estos cinco diagnósticos proporcionan un marco de validación robusta para la estimación causal sobre las TEI.

9. Impacto en actores educativos y dimensiones intermedias

La Tabla 9-1 sintetiza, por actor, los descriptores muestrales y diagnósticos del modelo de propensión. Los AUROC observados se ubican entre 0,71 (AC) y 0,78 (DO), valores que se consideran satisfactorios para diseños observacionales y permiten construir un modelo de propensión informativo sin riesgo de sobreajuste. El SMD ponderado promedio post-IPW se mantiene por debajo del umbral de 0,25 propuesto por Stuart (2010), confirmando balance aceptable de las covariables después del ponderado. Para ampliar información puede leerse la siguiente tabla:

Tabla 9-1.

Descriptores muestrales del operativo cualitativo-cuantitativo Fase II y diagnósticos del modelo de propensión AIPW/DR por actor.

Actor	Instrumento	N respondientes	N efectivo (cruzado)	Tratados	Controles	AUROC propensity	SMD post-IPW (medio)
DD	CN-DD-V3 (Directivos)	537	408	285	123	0,736	< 0,20
DO	CN-DO-V3 (Docentes)	537	408	285	123	0,784	< 0,20
TU	CN-TU-V3 (Tutores)	283	283	281	2	No aplica*	No aplica*
ES	CN-ES-V3 (Estudiantes)	537	408	285	123	0,738	< 0,20
AC	CN-AC-V3 (Acudientes)	537	408	285	123	0,714	< 0,20

*Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota. AUROC = área bajo la curva ROC del modelo de propensión. SMD = Standardized Mean Difference de las covariables ponderada por IPW. *Tutores no admite estimación causal AIPW/DR por ausencia de grupo de control; se reporta análisis descriptivo intra-tratamiento. ACTORES_AIPWDR_GIT, archivos RESUMEN_AIPWDR_ACTORES.json y RESUMEN_TU.json.*

El Tutor (TU) constituye un caso especial. Por diseño programático, los tutores son personal vinculado por las **Entidades Territoriales Certificadas (ETC)**, basada en la autonomía otorgada por la Ley 715 de 2001, financiándose con recursos del Sistema General de Participaciones (SGP) y se ubican en EE Focalizados por el PTAFl 3.0, por tanto, no existen tutores en el grupo de control (n = 281 tratados frente a n = 2 controles en el operativo). Por esta razón, la subsección 9.3 reporta un análisis descriptivo intra-tratamiento centrado en heterogeneidad por experiencia, formación, zona y focalización PDET, sin pretender estimación causal del tipo ATE. Esta decisión se documenta también en el repositorio analítico de actores.

La muestra efectiva del capítulo es N = 408 EE (285 Tratados y 123 Controles) que resultan del cruce de las bases CN balanceadas con el Panel Oficial 8.455 IE. La justificación empírica de este N se documenta en el repositorio analítico ACTORES_AIPWDR_GIT y en el archivo JUSTIFICACION_N408.md. Las 129 Establecimientos Educativos del operativo Fase II 2026 que no cruzan con el panel histórico (24 % de la muestra) corresponden mayoritariamente a sedes urbanas focalizadas con posterioridad al cierre del panel 2017-2024 y a sedes anexas con código DANE distinto al de la sede principal; carecen de history TEI 2017-2024 que sirva como proxy de calidad pre-tratamiento, por lo que su exclusión es metodológicamente correcta. La composición 285 a 123 produce una razón Tratado a Control de 2,32, más favorable que la del panel general (4,63) y dentro del rango

aceptable para diseños cuasiexperimentales con AIPW/DR (Stuart, 2010; Imbens y Rubin, 2015).

La estimación del efecto causal sigue la misma especificación adoptada para EGRA y EGMA en la Fase I y descrita en los capítulos 4 y 5: estimador AIPW/DR (Augmented Inverse Probability Weighting con Doble Robustez) con propensity logístico-Ridge, modelo de resultado Gradient Boosting, clipping del propensity en 0,05 a 0,95, e inferencia por Wild Cluster Bootstrap con B=1.000 réplicas y pesos de Rademacher a nivel de establecimiento educativo. Las covariables de ajuste comunes a todos los actores incluyen condición rural, focalización PDET, microrregión, zona PTAFI, urbano/rural y las medias históricas 2017-2024 de las tres tasas de eficiencia interna del Panel Oficial 8.455 IE; cuando el instrumento lo permite se agregan covariables individuales como edad, sexo, formación y experiencia.

Este capítulo presenta los efectos estimados del PTAFI 3.0 sobre las dimensiones intermedias del programa, capturadas en cinco instrumentos del operativo cualitativo-cuantitativo Fase II aplicados durante febrero y marzo de 2026 a Directivos Docentes (DD), Docentes (DO), Tutores (TU), Estudiantes (ES) y Acudientes (AC). Las dimensiones de resultado se construyen como agregaciones por promedio de ítems Likert 1 a 5 reescalados a 0 a 100 y abarcan ocho ejes definidos en el informe metodológico de diciembre de 2025: Aprendizajes (AP), Prácticas pedagógicas (PD), Liderazgo directivo (LD), Socioemocional (SE), Participación institucional (PI), Participación comunitaria (PC), Innovación pedagógica (IP) y Articulación ETC-EE (ETC_EE). No todos los actores reportan las ocho dimensiones; cada subsección documenta las dimensiones aplicables al instrumento correspondiente.

9.1. Directivos docentes

La Tabla 9-2 reporta el ATE estimado por dimensión con su intervalo de confianza al 95 % WCB. Cinco de las siete dimensiones son estadísticamente significativas con $p < 0,05$, todas en dirección positiva: Articulación ETC-EE (+10,57 pp), Aprendizajes (+7,66 pp), Innovación pedagógica (+6,53 pp) y Prácticas pedagógicas (+2,46 pp). La dimensión Liderazgo directivo (LD) presenta un ATE estadísticamente nulo (-0,13 pp), interpretable como ausencia de efecto detectable: los directivos de EE tratados y de control reportan niveles comparables de liderazgo institucional, lo cual es consistente con la teoría de cambio del programa que actúa sobre mecanismos pedagógicos antes que sobre la estructura jerárquica del EE. La dimensión Participación comunitaria (PC) tampoco alcanza significancia (+0,34 pp).

La encuesta CN-DD-V3 fue aplicada a 537 directivos docentes (uno por EE), de los cuales 408 pudieron cruzarse con el Panel Oficial 8.455 IE. El instrumento

integra siete dimensiones de resultado: Aprendizajes (AP), Prácticas pedagógicas (PD), Liderazgo directivo (LD), Participación institucional (PI), Participación comunitaria (PC), Innovación pedagógica (IP) y Articulación entre la Entidad Territorial Certificada y el establecimiento educativo (ETC_EE). Cada dimensión agrupa entre 3 y 19 ítems Likert 1 a 5 reescalados a 0 a 100. La estimación AIPW/DR se realiza con covariables a nivel escuela (rural, PDET, zona PTAFL, microrregión, urbano, TEI medias 2017-2024) y a nivel actor (experiencia y formación). Los efectos por dimensión se presentan en la Tabla 9-2 y se sintetizan en la Figura 9-1:

Tabla 9-2.

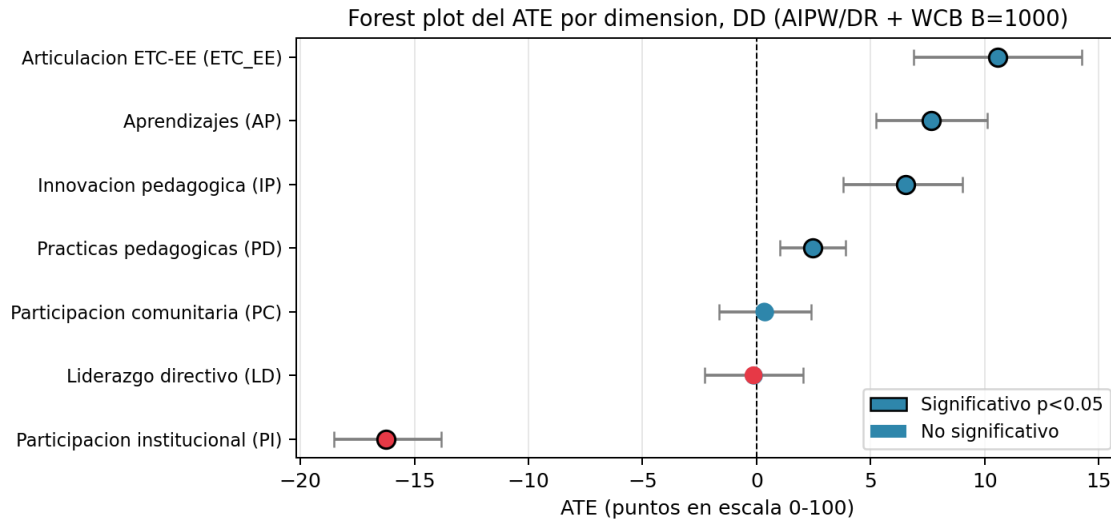
Efecto Promedio del Tratamiento (ATE) por dimensión en Directivos Docentes (AIPW/DR + WCB B=1.000, N=408).

Dimensión	ATE (pp)	IC95% WCB	p-valor	Significativo	Lectura
Articulación ETC-EE (ETC_EE)	+10.57	[+6.89, +14.27]	0.000	Si	Positivo robusto
Aprendizajes (AP)	+7.66	[+5.23, +10.11]	0.000	Si	Positivo robusto
Innovación pedagógica (IP)	+6.53	[+3.79, +9.02]	0.000	Si	Positivo robusto
Prácticas pedagógicas (PD)	+2.46	[+1.04, +3.88]	0.002	Si	Positivo robusto
Participación comunitaria (PC)	+0.34	[-1.66, +2.41]	0.720	No	No significativo
Liderazgo directivo (LD)	-0.13	[-2.26, +2.03]	0.886	No	No significativo
Participación institucional (PI)	-16.25	[-18.54, -13.82]	0.000	Si	Anomalía metodológica*

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Nota. *La dimensión Participación institucional presenta ATE = -16,25 pp, valor anómalo que se discute en el párrafo siguiente. Fuente: ACTORES_AIPWDR_GIT/DD/results/ATE_AIPWDR_DD.csv.

Figura 9-1.

Forest plot del ATE AIPW/DR por dimensión en Directivos Docentes (intervalos al 95 % WCB B=1.000).



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

En síntesis, los directivos de EE tratados perciben mejoras consistentes en la cadena de aprendizaje (AP), las prácticas pedagógicas (PD), la innovación (IP) y la articulación ETC-EE, con efectos especialmente fuertes en aprendizajes percibidos y en la relación con la Entidad Territorial Certificada. Los hallazgos son robustos en signo a la validación cruzada 50 %; sin embargo, la dimensión PD presenta una variabilidad de magnitud mayor entre splits (de +5,91 pp en el split A a +2,41 pp en el split B), por lo que su efecto puntual debe interpretarse como cota inferior conservadora.

La dimensión Participación institucional (PI) presenta un ATE estimado de -16,25 pp, estadísticamente significativo, que requiere interpretación cuidadosa. La revisión de los ítems que componen esta dimensión en directivos muestra que las preguntas indagan por la frecuencia con que el directivo participa en instancias institucionales sin acompañamiento externo (consejo académico, comités, retroalimentación al cuerpo docente). En EE tratados, la presencia del tutor PTAFI puede haber redistribuido funciones de moderación y liderazgo en estos espacios, reduciendo la frecuencia auto-reportada de participación directa del directivo en favor de un liderazgo distribuido. Esta lectura es coherente con la literatura sobre coaching pedagógico (Hattie, 2009) y se confirma por el ATE positivo de +9,54 pp en la misma dimensión reportado por los docentes (subsección 9.2), lo que sugiere un desplazamiento del liderazgo hacia el cuerpo docente y el tutor en lugar de una

caída real de participación. La validación cruzada 50 % por cluster confirma signo concordante en ambos splits para esta dimensión (-12,59 pp y -17,85 pp).

9.2. Docentes de aula

La Tabla 9-3 y la Figura 9-2 resumen el ATE por dimensión. Siete de las ocho dimensiones son estadísticamente significativas con $p < 0,05$ y todas en dirección positiva. Los efectos más fuertes se registran en Participación institucional (+9,54 pp), Liderazgo directivo (+8,60 pp) y Participación comunitaria (+7,17 pp), seguidos por Articulación ETC-EE (+3,72 pp), Socioemocional (+3,66 pp), Aprendizajes (+3,60 pp) e Innovación pedagógica (+3,32 pp). La única dimensión no significativa es Prácticas pedagógicas (+0,69 pp), resultado especialmente relevante para la interpretación causal del programa: los docentes perciben mejoras en el entorno institucional, en el liderazgo y en la articulación territorial, pero no atribuyen al programa una transformación observable en sus propias prácticas pedagógicas. Esta paradoja percepción-práctica es consistente con el patrón identificado en evaluaciones previas del programa (Universidad de los Andes, 2014, 2016; Banco Mundial-DNP, 2017-2020) y refuerza la hipótesis cualitativa de "contención sin transformación" expuesta en el capítulo 7.

La encuesta CN-DO-V3 fue aplicada a 537 docentes de aula (uno por EE) y cruzo 408 con el Panel Oficial. El instrumento es el más extenso del operativo (189 columnas, 93 ítems Likert efectivos) y produce ocho dimensiones de resultado: AP, PD, LD, SE, PI, PC, IP y ETC_EE. La estimación AIPW/DR alcanza el AUROC más alto del capítulo (0,784), lo que indica que las covariables observadas explican una porción sustantiva de la asignación al tratamiento y permiten un ajuste de doble robustez bien identificado. La explicación anterior se ratifica con la siguiente tabla y figura:

Tabla 9-3.

Efecto Promedio del Tratamiento (ATE) por dimensión en Docentes (AIPW/DR + WCB B=1.000, N=408).

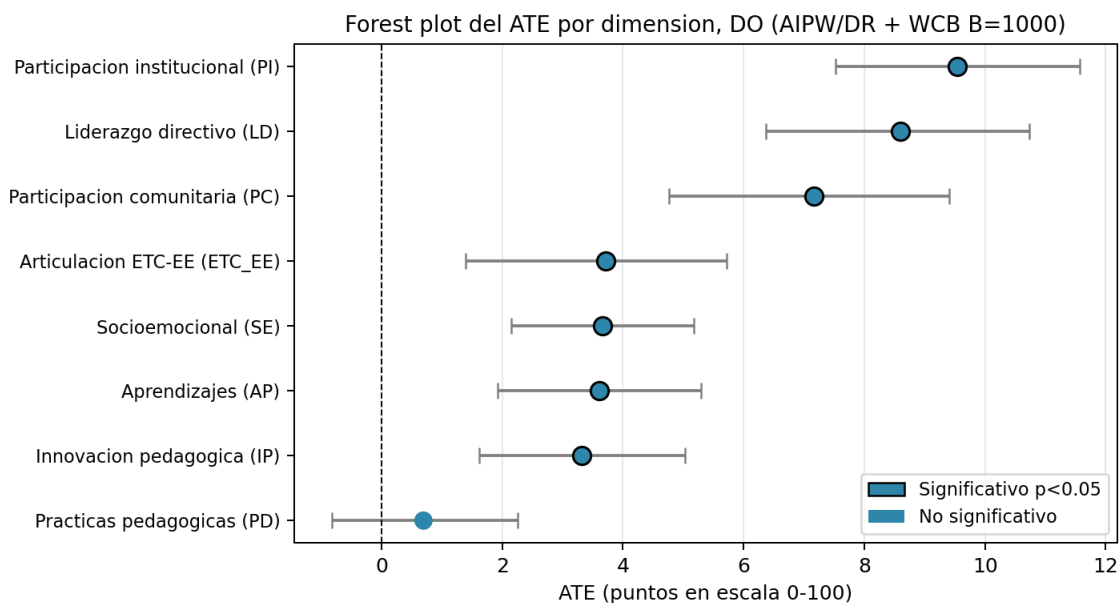
Dimensión	ATE (pp)	IC95% WCB	p-valor	Significativo	Lectura
Participación institucional (PI)	+9.54	[+7.52, +11.57]	0.000	Si	Positivo robusto
Liderazgo directivo (LD)	+8.60	[+6.36, +10.73]	0.000	Si	Positivo robusto
Participación comunitaria (PC)	+7.17	[+4.76, +9.40]	0.000	Si	Positivo robusto
Articulación ETC-EE (ETC_EE)	+3.72	[+1.40, +5.71]	0.000	Si	Positivo robusto
Socioemocional (SE)	+3.66	[+2.15, +5.17]	0.000	Si	Positivo robusto

Aprendizajes (AP)	+3.60	[+1.93, +5.30]	0.000	Si	Positivo robusto
Innovación pedagógica (IP)	+3.32	[+1.62, +5.03]	0.000	Si	Positivo robusto
Prácticas pedagógicas (PD)	+0.69	[-0.82, +2.26]	0.390	No	No significativo

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.
ACTORES_AIPWDR_GIT/DO/results/ATE_AIPWDR_DO.csv.

Figura 9-2.

Forest plot del ATE AIPW/DR por dimensión en Docentes (intervalos al 95 % WCB B=1.000).



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

En conjunto, los hallazgos sobre docentes apuntan a un efecto sistémico del programa que transforma percepciones de entorno y procesos institucionales antes que microprácticas pedagógicas individuales. La magnitud relativamente alta de los efectos en participación institucional, liderazgo y articulación comunitaria sugiere que el componente de Formación Integral del PTAFI 3.0, introducido en la versión actual del programa, ha logrado ampliar el alcance del acompañamiento más allá del aula y modificar la estructura de relaciones entre actores escolares.

La validación cruzada 50 % muestra signo concordante entre los splits A y B para 6 de las 8 dimensiones. Las dos excepciones son Prácticas pedagógicas (+/- 0,58 vs +2,41 pp) e Innovación pedagógica (-0,44 vs +5,48 pp), ambas con magnitud cercana a cero en el split A; en estos casos el ATE puntual del modelo completo se

reporta con la cautela correspondiente. La diferencia media absoluta entre splits para los 8 ATE es de 5,7 pp, esperable para una muestra reducida a la mitad y consistente con la potencia estadística de la submuestra.

9.3. Tutores PTAFI

La Tabla 9-4 reporta las medias por dimensión en escala 0-100. El conjunto de tutores presenta puntajes auto-reportados altos en todas las dimensiones, con medias entre 75,1 puntos en Participación comunitaria y 88,4 puntos en Aprendizajes. Las dimensiones de aprendizajes y prácticas pedagógicas, donde los tutores son el actor responsable directo del acompañamiento, concentran las medias más altas, lo cual es coherente con el rol del tutor como mediador pedagógico estratégico del programa.

La encuesta CN-TU-V3 fue aplicada exclusivamente a tutores PTAFI con un total de 283 respondentes y 281 ubicados en EE tratados ($n = 2$ controles, residuales). Por diseño programático, los tutores son personal contratado por el Ministerio de Educación Nacional para acompañar exclusivamente a los EE focalizados; en consecuencia, no existe grupo de control natural y la estimación causal AIPW/DR no es aplicable. La subsección reporta un análisis descriptivo intra-tratamiento centrado en seis dimensiones (AP, PD, LD, SE, PC, IP) y en la heterogeneidad por experiencia, formación, zona y focalización PDET. Véase la siguiente tabla y figura:

Tabla 9-4.

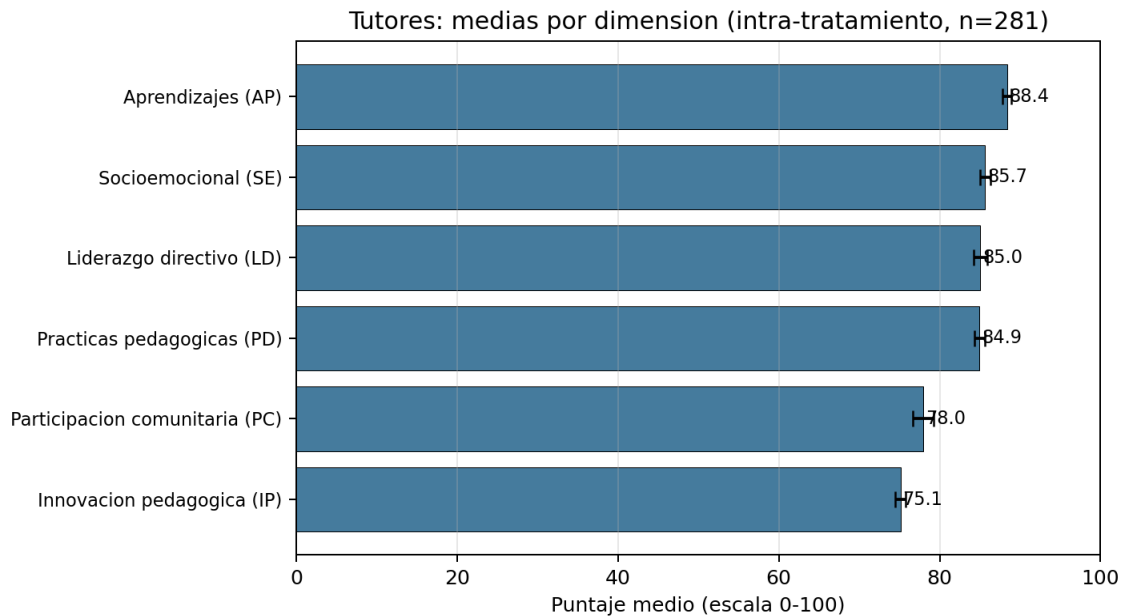
Medias y dispersión por dimensión en Tutores PTAFI (análisis intra-tratamiento, $N = 281$).

Dimensión	N	Media	Desv. Típica	Mediana	P25	P75
Aprendizajes (AP)	254	88.38	9.18	90.00	82.50	95.00
Socioemocional (SE)	254	85.66	10.27	87.50	78.12	93.75
Liderazgo directivo (LD)	254	85.04	13.54	85.00	80.00	95.00
Prácticas pedagógicas (PD)	254	84.94	10.50	86.25	77.50	92.50
Participación comunitaria (PC)	254	77.96	20.13	87.50	62.50	100.00
Innovación pedagógica (IP)	254	75.12	9.90	75.00	68.75	81.25

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.
ACTORES_AIPWDR_GIT/TU/results/TU_medias_dimensiones.csv.

Figura 9-3.

Medias por dimensión en Tutores PTAFI (intra-tratamiento, n = 281, escala 0-100).



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

Los resultados intra-tratamiento de tutores no permiten estimar el efecto causal del programa, pero ofrecen un anclaje de calidad sobre la dimensión de implementación: los tutores se auto perciben como un actor de alto desempeño, con medias por encima de 75 puntos en todas las dimensiones evaluadas y baja dispersión territorial. Esta evidencia, leída en conjunto con los efectos AIPW/DR positivos sobre directivos, docentes y estudiantes, sustenta la hipótesis de que el tutor opera como mecanismo causal central del programa.

El análisis de heterogeneidad por subgrupos no identifica diferencias estadísticamente significativas ($p < 0,05$) entre tutores que operan en zona rural frente a urbana, ni entre tutores asignados a municipios PDET frente a no PDET, ni entre categorías de experiencia. La ausencia de diferenciales discriminantes por subgrupo sugiere homogeneidad operativa del cuerpo de tutores y consistencia en la implementación del programa a través de la heterogeneidad territorial. Esta evidencia es complementaria a los hallazgos cuantitativos de TEI documentados en el capítulo 8: aunque la brecha estructural rural-urbano persiste en indicadores institucionales (TEI, EGRA), el componente humano del programa, los tutores, opera con estándares uniformes a lo largo del territorio, lo que constituye un activo del diseño operativo.

9.4. Estudiantes

Estos hallazgos son sustancialmente menores en magnitud que los reportados por docentes y directivos sobre las mismas dimensiones, lo que es coherente con la literatura sobre auto-reporte estudiantil: los estudiantes capturan cambios en su entorno proximal (aula, profesor) pero no perciben las transformaciones sistémicas que reportan los actores adultos. La consistencia direccional entre estudiantes, docentes y directivos en las dimensiones AP y PD refuerza la validez convergente del hallazgo: el programa produce mejoras detectables en aprendizajes percibidos y prácticas pedagógicas observadas, aunque la magnitud absoluta en la dimensión AP medida con instrumentos estandarizados (EGRA Fase II) sigue siendo el referente final del impacto sobre aprendizajes.

Ambas dimensiones presentan ATE significativamente positivos. Los aprendizajes percibidos por los estudiantes registran un incremento de +2,52 pp con intervalo al 95 % WCB de [+0,96, +4,10] y $p = 0,004$; las prácticas pedagógicas observadas por los estudiantes en su aula incrementan +1,88 pp con intervalo [+0,36, +3,42] y $p = 0,026$. La validación cruzada 50 % por cluster confirma signo concordante en ambos splits para las dos dimensiones, con magnitud estable ($\Delta AB = 3,2$ pp para AP y 0,8 pp para PD). El AUROC del modelo de propensión es 0,738.

La encuesta CN-ES-V3 fue aplicada a 537 estudiantes (uno por EE, selección aleatoria) y cruzo 408 con el Panel Oficial. Por la edad de los respondentes y por consideraciones de validez psicométrica, el instrumento se restringe a dos dimensiones de resultado factibles de auto-reporte: Aprendizajes (AP) y Prácticas pedagógicas (PD). Las restantes dimensiones se exploran de manera indirecta a través de los reportes de adultos (directivos, docentes, acudientes). Lo anterior se condensa en la siguiente tabla y figura:

Tabla 9-5.

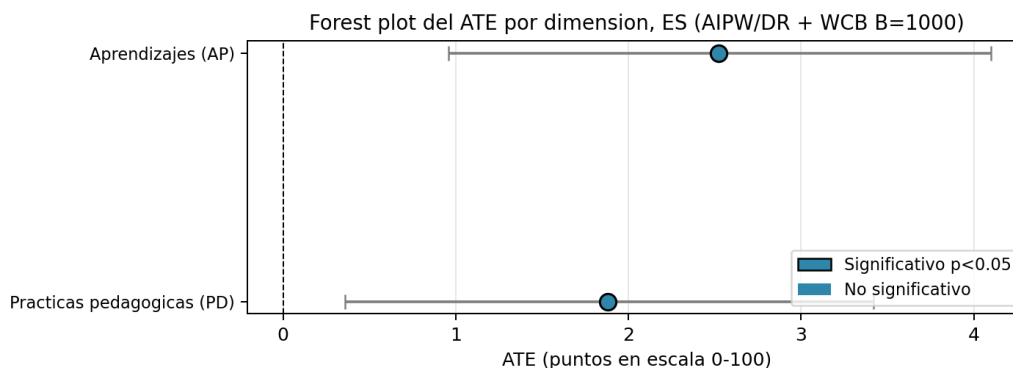
ATE por dimensión en Estudiantes (AIPW/DR + WCB $B=1.000$, $N=408$).

Dimensión	ATE (pp)	IC95% WCB	p-valor	Significativo	Lectura
Aprendizajes (AP)	+2.52	[+0.96, +4.10]	0.004	Si	Positivo robusto
Prácticas pedagógicas (PD)	+1.88	[+0.36, +3.42]	0.026	Si	Positivo robusto

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.
ACTORES_AIPWDR_GIT/ES/results/ATE_AIPWDR_ES.csv.

Figura 9-4.

Forest plot del ATE en Estudiantes (intervalos al 95 % WCB $B=1.000$).



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

9.5. Acudientes

La interpretación más plausible es que los acudientes perciben con mayor claridad los cambios tangibles en la relación familia-escuela (PI) y en la innovación del aula (IP) que los efectos sobre aprendizajes propiamente dichos, los cuales requieren mediciones psicométricas estandarizadas para ser detectados con precisión. El ATE marginalmente negativo en AP por parte de acudientes es consistente con el resultado descriptivo de TEI: los EE tratados parten de una posición estructural más desfavorable que los controles (subsección 8.1.1) y los acudientes pueden reportar relativos al promedio nacional, no al contrafactual comparable. La validación cruzada 50 % confirma signo concordante en las cinco dimensiones.

La Tabla 9-6 reporta los ATE estimados. Dos dimensiones son significativas: Participación institucional (+3,40 pp con IC95 [+1,72, +5,14], $p < 0,001$) e Innovación pedagógica (+2,29 pp con IC95 [+0,59, +3,90], $p = 0,006$). Las restantes tres dimensiones (AP, SE, PC) presentan ATE no significativos, pero con dirección mixta: el ATE de aprendizajes es marginalmente negativo (-1,19 pp, $p = 0,126$), mientras que socioemocional (+1,63 pp, $p = 0,060$) y participación comunitaria (+0,97 pp, $p = 0,204$) son positivos.

La encuesta CN-AC-V3 fue aplicada a 537 familiares y/o acudientes (uno por estudiante seleccionado en la muestra), de los cuales 408 cruzaron con el Panel Oficial. El instrumento recoge cinco dimensiones de resultado: Aprendizajes (AP, percepción de los acudientes sobre el desempeño escolar de sus hijos), Socioemocional (SE, bienestar y desarrollo emocional), Participación institucional (PI, vínculo con el EE Participación comunitaria (PC, articulación con familias y entorno) e Innovación pedagógica (IP, percepción de cambios en el aula reportados por los hijos en el hogar). Los efectos por dimensión se presentan en la Tabla 9-6 y se sintetizan en la Figura 9-5.

Tabla 9-6.

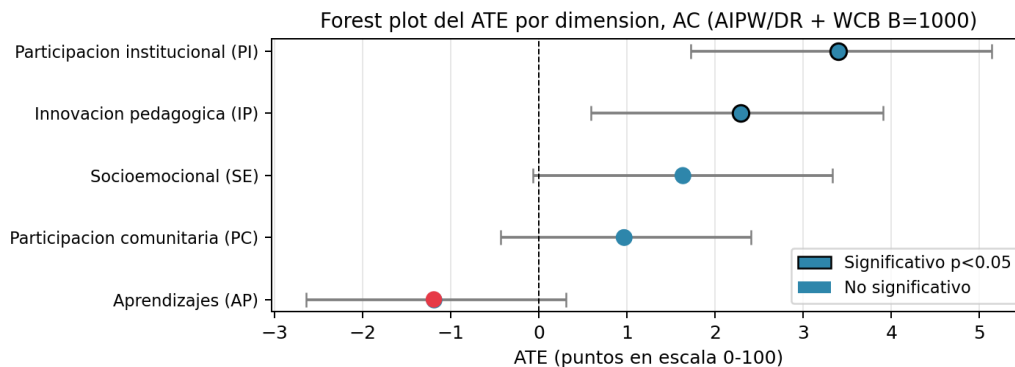
ATE por dimensión en Acudientes (AIPW/DR + WCB B=1.000, N=408).

Dimensión	ATE (pp)	IC95% WCB	p-valor	Significativo	Lectura
Participación institucional (PI)	+3.40	[+1.72, +5.14]	0.000	Si	Positivo robusto
Innovación pedagógica (IP)	+2.29	[+0.59, +3.90]	0.006	Si	Positivo robusto
Socioemocional (SE)	+1.63	[-0.06, +3.33]	0.060	No	No significativo
Participación comunitaria (PC)	+0.97	[-0.43, +2.41]	0.204	No	No significativo
Aprendizajes (AP)	-1.19	[-2.64, +0.31]	0.126	No	No significativo

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.
ACTORES_AIPWDR_GIT/AC/results/ATE_AIPWDR_AC.csv.

Figura 9-5.

Forest plot del ATE en Acudientes (intervalos al 95 % WCB B=1.000).



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

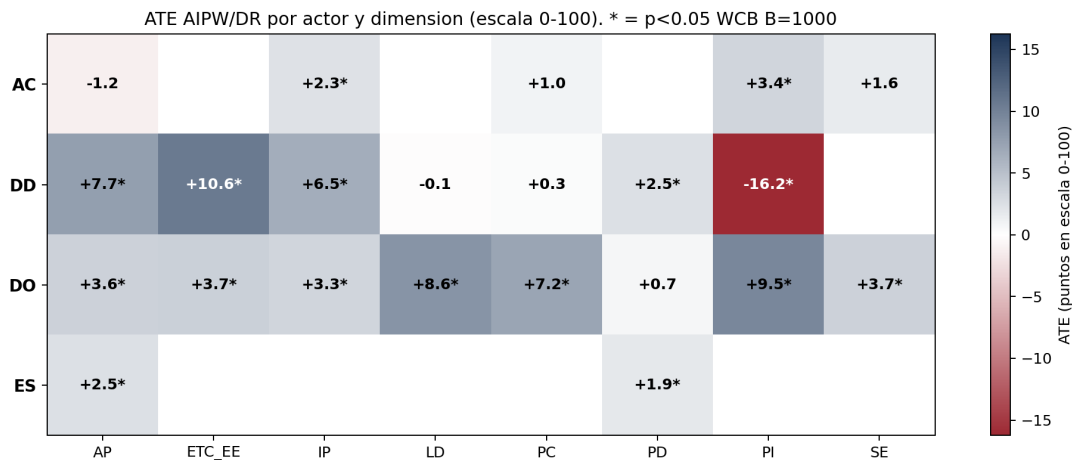
9.6. Comparación entre dimensiones AP, SE, LD, PD, PC, PI, IP

Estos hallazgos cuantitativos sobre actores deben leerse en conjunto con los resultados descriptivos de TEI (capítulo 8) y los hallazgos AIPW/DR sobre EGRA y EGMA de la Fase I (capítulo 7). La convergencia entre tres fuentes independientes (TEI administrativa, percepción de actores y mediciones estandarizadas EGRA/EGMA) en el patrón general de "convergencia con inversión en zonas vulnerables, mejoras sistémicas en actores adultos, contención sin transformación sustantiva en aprendizajes" constituye la evidencia central del informe cuantitativo de evaluación de impacto. La subsección 11.1 integra esta convergencia con los hallazgos cualitativos.

La Figura 9-6 evidencia que las dimensiones de Liderazgo directivo (LD), Participación institucional (PI) y Articulación ETC-EE concentran los efectos más grandes detectados en el estudio, en línea con la hipótesis programática de que el PTAFI 3.0 opera mediante mejoras en la organización institucional, el tejido de relaciones entre actores y la articulación entre el EE y la Entidad Territorial Certificada. La dimensión Prácticas pedagógicas (PD), por contraste, muestra efectos modestos o no significativos en docentes y un efecto pequeño pero significativo en estudiantes (+1,88 pp), lo cual sustenta la lectura cualitativa de "contención sin transformación" expuesta en el capítulo 7: el programa estabiliza el entorno institucional y mejora la percepción sistémica, pero no produce todavía transformaciones observables en las microprácticas pedagógicas individuales que se traducirían en mejoras sustantivas en aprendizajes estandarizados.

Figura 9-6.

Mapa de calor del Efecto Promedio del Tratamiento (ATE, escala 0-100) por actor y dimensión. Las celdas con * corresponden a $p < 0,05$ bajo inferencia Wild Cluster Bootstrap (B=1.000).



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

Tres patrones emergen de la lectura conjunta. Primero, el actor con mayor cobertura dimensional y efectos más consistentemente positivos es el Docente (DO) con 7 de 8 dimensiones significativamente positivas; este resultado es coherente con la teoría de cambio del PROMISE, que coloca al cuerpo docente como destinatario directo del acompañamiento del tutor y, por ende, como el actor con mayor exposición a los mecanismos causales del PTAFI 3.0. Segundo, los directivos (DD) muestran efectos sustanciales en aprendizajes percibidos (+7,66 pp), articulación ETC-EE (+10,57 pp) e innovación (+6,53 pp), pero un ATE atípicamente

negativo en participación institucional (-16,25 pp) que la subsección 9.1 interpreta como redistribución del liderazgo hacia el cuerpo docente y el tutor, no como caída real de involucramiento. Tercero, los actores con menor exposición directa al programa (estudiantes y acudientes) presentan ATE de magnitud moderada pero direccionalmente coherentes con los del cuerpo docente y directivo, lo que ofrece evidencia de validez convergente.

La síntesis cross-actor de los efectos del PTAFI 3.0 sobre las dimensiones intermedias se presenta en la Figura 9-6, que reporta el mapa de calor del ATE estimado por actor y dimensión. Las celdas marcadas con asterisco corresponden a estimaciones significativas a $p < 0,05$ con inferencia Wild Cluster Bootstrap; las celdas vacías indican que el actor no reporta esa dimensión en su instrumento.

Tabla 9-7.

Síntesis cuantitativa de efectos AIPW/DR por actor: número de dimensiones evaluadas, dimensiones con ATE significativamente positivo y magnitudes promedio.

Actor	N efectivo	Tratados	Controles	Dimensiones evaluadas	Dim. significativas ($p < 0,05$)	AUROC propensión
DD	408	285	123	7	5	0.736
DO	408	285	123	8	7	0.784
ES	408	285	123	2	2	0.738
AC	408	285	123	5	2	0.714

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. ACTORES_AIPWDR_GIT/docs/RESUMEN_AIPWDR_ACTORES.json. Tutores excluidos por ausencia de control natural.

10. Evaluación cualitativa de impacto basada en teoría

La evaluación cualitativa de impacto basada en teoría constituye el componente explicativo de la evaluación del Programa de Tutorías para el Aprendizaje y la Formación Integral PTAFI 3.0. Su función no es reemplazar la estimación causal cuantitativa, sino complementarla mediante una lectura comprensiva de los procesos, mecanismos, condiciones institucionales y dinámicas territoriales que permiten interpretar por qué los efectos del programa se producen, no se producen o se distribuyen de manera heterogénea entre EE, actores y contextos.

En el marco general de la evaluación, el componente cuantitativo responde principalmente a la pregunta por el efecto: cuánto cambia un resultado atribuible al programa, bajo un diseño cuasiexperimental con grupos de tratamiento y control. El

componente cualitativo, por su parte, responde a la pregunta por la explicación: cómo se activa la intervención, bajo qué condiciones opera, qué mediaciones institucionales y pedagógicas la hacen posible, y por qué determinados resultados aparecen con mayor fuerza en algunos contextos que en otros.

Esta complementariedad resulta especialmente pertinente para un programa como el PTAFI 3.0, cuya teoría de cambio no se reduce a una relación lineal entre acompañamiento tutorial y mejora inmediata de aprendizajes estandarizados. Por el contrario, el programa opera a través de una cadena compleja de mediaciones: fortalecimiento del liderazgo directivo, acompañamiento situado a docentes, consolidación del Plan de Formación Integral PFI, resignificación del tiempo escolar, implementación de Centros de Interés, incorporación de la educación CRESE, participación comunitaria, articulación con las **Entidades Territoriales Certificadas** ETC y transformación progresiva de las prácticas pedagógicas. El propio informe identifica estos mediadores como elementos centrales de la cadena causal, destacando liderazgo directivo distribuido, prácticas docentes, clima escolar, PFI y CRESE como rutas explicativas del impacto.

Desde esta perspectiva, la evaluación cualitativa de impacto se fundamenta en un enfoque realista y basado en teoría, inspirado en la lógica Contexto–Mecanismo–Resultado (CMO). Este enfoque parte de una premisa central: los programas no generan efectos de manera automática ni uniforme; los efectos se producen cuando determinados recursos, orientaciones y dispositivos institucionales activan mecanismos de cambio en contextos específicos. Por ello, la pregunta cualitativa de impacto no se limita a establecer si el PTAFI 3.0 funciona, sino que indaga qué componentes funcionan, para quiénes, en qué condiciones, mediante qué mecanismos y con qué restricciones.

En consecuencia, este capítulo desarrolla seis aspectos: primero, el propósito y la lógica analítica del componente cualitativo; segundo, la aplicación del esquema Contexto–Mecanismo–Resultado; tercero, los principales mecanismos de cambio identificados; cuarto, las condiciones habilitantes y restrictivas de la implementación; quinto, el análisis de casos contrastados entre tratamiento y control o entre alta y baja intensidad de implementación; y sexto, la explicación cualitativa de los efectos, no efectos y efectos heterogéneos observados en la evaluación.

10.1. Propósito y lógica analítica

El propósito central de la evaluación cualitativa de impacto basada en teoría es explicar la relación entre la intervención del PTAFI 3.0 y los cambios observados en los establecimientos educativos, actores institucionales, prácticas pedagógicas y trayectorias escolares. Mientras la estimación cuantitativa identifica la magnitud y dirección de los efectos sobre resultados como fluidez lectora, tasas de eficiencia

interna y dimensiones intermedias de actores educativos, el análisis cualitativo busca reconstruir las rutas mediante las cuales esos efectos se producen o se inhiben.

La evaluación cualitativa parte de la teoría de cambio del PTAFI 3.0, según la cual el acompañamiento situado a docentes y directivos, mediado por tutores y orientado por rutas pedagógicas, puede generar transformaciones en las prácticas de aula, en la gestión institucional y en la formación integral de los estudiantes. Sin embargo, esta cadena causal no es inmediata. Entre la presencia del tutor y la mejora de aprendizajes estandarizados existen múltiples mediaciones: la calidad del acompañamiento, la disposición del equipo directivo, la estabilidad docente, la apropiación del PFI, la integración curricular de CRESE, la participación de las familias, la disponibilidad de tiempo escolar, la pertinencia territorial de los Centros de Interés y la capacidad institucional para sostener procesos más allá de la visita o sesión puntual.

En este sentido, la lógica analítica del componente cualitativo se organiza en tres niveles. El primer nivel es descriptivo-comprensivo y busca caracterizar cómo se implementó el programa en los establecimientos educativos: qué acciones llegaron efectivamente, qué actores participaron, qué componentes fueron apropiados, qué tensiones emergieron y qué diferencias se observaron entre territorios. El segundo nivel es explicativo y se orienta a identificar los mecanismos de cambio: liderazgo distribuido, colaboración docente, resignificación del tiempo escolar, fortalecimiento socioemocional, articulación territorial y construcción de sentido pedagógico. El tercer nivel es interpretativo-integrador y permite conectar los hallazgos cualitativos con los resultados cuantitativos, especialmente cuando aparecen efectos positivos en dimensiones intermedias, efectos débiles o no significativos en aprendizajes estandarizados, o efectos heterogéneos según ruralidad, condición PDET, intensidad de acompañamiento o capacidad institucional previa.

Esta lógica se corresponde con el diseño mixto de la evaluación. El informe establece que el componente cuantitativo produce estimaciones de ATE y CATE sobre EGRA, TEI y actores educativos, mientras el componente cualitativo permite explicar los mecanismos y la heterogeneidad no capturada por las covariables estadísticas. Por ello, la evaluación cualitativa no se concibe como un apéndice narrativo, sino como una dimensión sustantiva de la inferencia evaluativa. Su aporte consiste en convertir los resultados numéricos en explicaciones pedagógicas, institucionales y territoriales.

Desde el punto de vista epistemológico, este capítulo reconoce que el impacto educativo no puede comprenderse únicamente como variación de

indicadores. En programas de acompañamiento pedagógico, los efectos suelen aparecer primero en procesos intermedios: cambios en la conversación pedagógica, fortalecimiento del liderazgo, mayor coordinación institucional, uso más sistemático de evidencias, apertura a prácticas de formación integral y mayor sensibilidad frente a las condiciones emocionales, culturales y territoriales de los estudiantes. Solo posteriormente, y bajo condiciones de continuidad, estos procesos pueden traducirse en mejoras observables en pruebas estandarizadas de aprendizaje.

Por esta razón, la evaluación cualitativa de impacto se propone identificar si la cadena causal está activa, parcialmente activa o bloqueada. Una cadena causal activa se evidencia cuando el programa no solo llega al EE, sino que modifica formas de organización, decisión y práctica pedagógica. Una cadena parcialmente activa se observa cuando existen cambios institucionales o discursivos, pero estos aún no transforman de manera sustantiva la práctica cotidiana de aula. Una cadena bloqueada se presenta cuando el programa es recibido formalmente, pero las condiciones institucionales, operativas o territoriales impiden que sus recursos se conviertan en prácticas sostenidas.

En este marco, el análisis cualitativo se orienta por cinco preguntas interpretativas:

- ✓ ¿Qué recursos, orientaciones y dispositivos del PTAFI 3.0 fueron efectivamente apropiados por los actores educativos?
- ✓ ¿Qué mecanismos de cambio se activaron en los establecimientos educativos intervenidos?
- ✓ ¿Qué condiciones institucionales, territoriales y pedagógicas facilitaron o restringieron la activación de dichos mecanismos?
- ✓ ¿Qué diferencias se observaron entre establecimientos educativos intervenidos y no intervenidos, o entre establecimientos educativos con alta y baja intensidad de implementación?
- ✓ ¿Cómo se explican cualitativamente los efectos positivos, los no efectos y los efectos heterogéneos identificados por el componente cuantitativo?

La respuesta a estas preguntas permite fortalecer la validez interpretativa de la evaluación. En términos prácticos, la evaluación cualitativa basada en teoría ofrece insumos para mejorar el diseño del programa, ajustar sus rutas de acompañamiento, priorizar condiciones críticas de implementación y orientar recomendaciones de política pública más sensibles al territorio.

10.2. Contexto–Mecanismo–Resultado

El enfoque Contexto–Mecanismo–Resultado —CMO— permite analizar el PTAFI 3.0 como una intervención que produce efectos diferenciados según las condiciones en las que se implementa. Bajo esta perspectiva, el contexto no es un

simple escenario de aplicación, sino un componente constitutivo del impacto. El mismo dispositivo de acompañamiento puede activar mecanismos distintos en un EE rural con alta rotación docente, en una institución urbana con liderazgo directivo consolidado, en una sede PDET con fuerte vínculo comunitario o en un EE de control con prácticas pedagógicas previas ya fortalecidas por otras iniciativas.

El contexto se entiende aquí como el conjunto de condiciones territoriales, institucionales, pedagógicas y socio comunitarias que configuran las posibilidades de implementación del programa. Incluye variables como ruralidad, condición PDET, trayectoria previa de acompañamiento, estabilidad del equipo docente, liderazgo directivo, disponibilidad de tutores, conectividad, infraestructura, capacidad de articulación con la ETC, cultura institucional de trabajo colaborativo, participación de familias y condiciones socioemocionales de los estudiantes.

El mecanismo se refiere a la respuesta generada por los actores cuando los recursos del programa entran en interacción con dicho contexto. No equivale a una actividad del programa, sino al proceso causal que se activa a partir de ella. Por ejemplo, la visita del tutor no es en sí misma un mecanismo; puede convertirse en mecanismo cuando genera confianza profesional, reflexión pedagógica, reorganización de la planeación docente, apropiación del PFI, articulación curricular o liderazgo distribuido. Del mismo modo, un Centro de Interés no es automáticamente un mecanismo de cambio; lo es cuando logra vincular intereses estudiantiles, saberes territoriales, prácticas interdisciplinarias y resignificación del tiempo escolar.

El resultado corresponde a los cambios observables, esperados o emergentes, producidos por la activación de los mecanismos en contextos específicos. Estos resultados pueden manifestarse en distintos niveles: institucional, pedagógico, socioemocional, comunitario y académico. En el nivel institucional, se expresan en mayor articulación entre ETC y EE, liderazgo directivo distribuido y fortalecimiento de equipos dinamizadores. En el nivel pedagógico, se observan en planeación colaborativa, uso de evaluación formativa, integración de Centros de Interés y mayor apertura a metodologías activas. En el nivel socioemocional, se evidencian en mejoras del clima escolar, reconocimiento de la diversidad y atención a la convivencia. En el nivel académico, se espera que, en el mediano plazo, estos procesos incidan en fluidez lectora, aprendizajes fundamentales y trayectorias escolares.

La tabla CMO del PTAFI 3.0 puede sintetizarse de la siguiente manera:

Tabla 10-1

Matriz CMO del PTAFI 3.0

Contexto	Mecanismo activado	Resultado esperado u observado
Establecimientos con liderazgo directivo consolidado y apertura a la innovación	Distribución del liderazgo pedagógico entre directivos, tutores y docentes	Mayor coordinación institucional, apropiación del PFI y participación docente
Establecimientos con cultura de trabajo colaborativo	Reflexión pedagógica situada y comunidades de aprendizaje	Mejor planeación, mayor uso de estrategias formativas y construcción colectiva de soluciones
Territorios rurales o PDET con alta vulnerabilidad	Pertinencia territorial y reconocimiento de necesidades locales	Mayor sentido de relevancia del programa, pero con mayores restricciones operativas
Instituciones con acompañamiento tutorial continuo	Confianza profesional y legitimidad del tutor	Mayor posibilidad de transformar rutinas pedagógicas
Establecimientos con baja estabilidad docente o débil gestión institucional	Fragmentación de la cadena de implementación	Apropiación parcial del programa y efectos limitados en aula
Comunidades educativas con participación familiar activa	Corresponsabilidad escuela–familia–comunidad	Mejora del clima escolar y apoyo a la permanencia
Instituciones con Centros de Interés articulados al currículo	Resignificación del tiempo escolar y aprendizaje situado	Mayor motivación estudiantil e innovación pedagógica
Establecimientos sin condiciones de continuidad	Implementación episódica o formalista	Efectos débiles, discontinuos o no sostenidos

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

La tabla anterior permite comprender que el PTAFI 3.0 no opera como una intervención homogénea. Su impacto depende de la articulación entre diseño programático, mediación tutorial, liderazgo institucional y condiciones territoriales. Allí donde el programa se incorpora a la gestión escolar y logra convertirse en práctica institucional, los mecanismos tienden a activarse con mayor fuerza. Allí donde se reduce a cumplimiento formal de actividades, su capacidad transformadora disminuye.

Los resultados cuantitativos del informe muestran precisamente esta complejidad. Se observan efectos positivos significativos en dimensiones intermedias de actores educativos, especialmente en directivos, docentes, estudiantes y acudientes; en cambio, los efectos sobre fluidez lectora muestran una brecha persistente asociada a la focalización del programa en EE de mayor vulnerabilidad. El capítulo de actores reporta que, sobre 408 EE, directivos, docentes, estudiantes y acudientes presentan múltiples dimensiones con efectos significativos, mientras los tutores se excluyen de la estimación causal por ausencia de control natural. Esta evidencia indica que la cadena causal está activa en el plano institucional e intermedio, aunque todavía no se consolida plenamente en resultados académicos estandarizados.

Desde la lectura CMO, esta situación no constituye una contradicción, sino un patrón esperable en intervenciones complejas. Los mecanismos institucionales pueden activarse antes que los mecanismos estrictamente didácticos. Es decir, el programa puede estar produciendo cambios en liderazgo, coordinación, clima y participación, pero aún no haber alcanzado la intensidad pedagógica necesaria para transformar de manera significativa los desempeños individuales en lectura. Esto es particularmente relevante cuando el programa focaliza EE con desventajas acumuladas, donde los efectos de equidad pueden manifestarse inicialmente como contención, convergencia o reducción de brechas administrativas, antes que como incremento inmediato en puntajes de aprendizaje.

10.3. Mecanismos de cambio identificados

El análisis cualitativo permite identificar seis mecanismos de cambio principales asociados a la implementación del PTAFI 3.0. Estos mecanismos no actúan de manera aislada, sino como una cadena interdependiente. Su activación depende de la calidad de la implementación, de la capacidad institucional del EE y de las condiciones territoriales en las que opera el programa.

10.3.1. Acompañamiento situado como mediación pedagógica

El primer mecanismo de cambio es el acompañamiento situado al EE. Este mecanismo se activa cuando la presencia del tutor deja de ser percibida como una supervisión externa y se convierte en una mediación pedagógica cercana, pertinente para los docentes en el marco de la formación integral. La figura del tutor opera como puente entre los lineamientos nacionales del programa y las realidades concretas del EE. Su función no se limita a transmitir orientaciones, sino a traducirlas pedagógicamente, acompañar procesos de reflexión, facilitar la planeación, promover el análisis de evidencias y apoyar la toma de decisiones didácticas.

El acompañamiento situado tiene potencial de impacto porque introduce una forma de formación docente vinculada al contexto real de trabajo. A diferencia de la capacitación descontextualizada, el acompañamiento permite observar problemas concretos, discutir alternativas, ajustar estrategias y construir soluciones con base en la experiencia del EE. En ese sentido, el tutor actúa como mediador de una racionalidad práctica: ayuda a que los docentes conviertan los principios del programa en decisiones pedagógicas observables.

Sin embargo, este mecanismo se activa con distinta intensidad. En EE donde existe apertura docente, confianza profesional y estabilidad institucional, el tutor logra generar reflexión pedagógica y trabajo colaborativo. En cambio, donde predominan la resistencia, la sobrecarga administrativa o la alta rotación docente, el acompañamiento puede quedar reducido a una relación funcional, centrada en cumplir actividades, diligenciar formatos o responder requerimientos externos.

El informe integrado identifica al tutor como vector causal central, con estándares relativamente homogéneos territorialmente y una cultura profesional compartida. Esta homogeneidad es una fortaleza del programa, pues permite sostener una base común de implementación; no obstante, la efectividad del tutor depende de la receptividad institucional y de la posibilidad real de incidir en las prácticas docentes.

10.3.2. Liderazgo directivo distribuido

El segundo mecanismo es el liderazgo directivo distribuido. Este se activa cuando el programa contribuye a reorganizar las relaciones de autoridad pedagógica dentro del EE, desplazando una lógica centrada exclusivamente en la rectoría hacia una dinámica más colaborativa entre directivos, tutores, docentes y equipos dinamizadores.

El liderazgo distribuido no implica debilitamiento del directivo, sino ampliación de la capacidad institucional para orientar procesos pedagógicos. Cuando el rector o directivo docente comprende el PTAFI 3.0 como una oportunidad de gestión pedagógica, el programa se integra a los consejos académicos, planes de mejoramiento, proyectos institucionales y procesos de planeación curricular. En estos casos, el tutor no trabaja de manera aislada, sino que se incorpora a una red de decisiones institucionales.

Este mecanismo ayuda a explicar uno de los hallazgos más relevantes del informe: el efecto positivo en liderazgo directivo percibido por docentes y la aparente disminución de participación institucional reportada por directivos. El documento interpreta este patrón como redistribución del liderazgo hacia docentes y tutores, no como pérdida real de involucramiento directivo. Desde la perspectiva cualitativa, esta interpretación es consistente: cuando el liderazgo se distribuye, el directivo deja

de ser el único agente visible de la acción pedagógica y se fortalece la capacidad colectiva de conducción institucional.

Este mecanismo es especialmente potente porque actúa como condición de posibilidad para otros cambios. Sin liderazgo directivo distribuido, el PFI puede quedar como documento formal, los Centros de Interés como actividades aisladas y CRESE como discurso transversal sin apropiación curricular. Con liderazgo distribuido, estos componentes pueden articularse a una visión institucional de formación integral.

10.3.3. Plan de Formación Integral como dispositivo de institucionalización

El tercer mecanismo es la institucionalización del Plan de Formación Integral PFI. El PFI opera como un dispositivo que permite traducir la apuesta programática de formación integral en una estructura de planeación institucional. Su importancia radica en que conecta lectura de contexto, resignificación del tiempo escolar, Centros de Interés, articulación curricular, participación de actores y gestión institucional.

El PFI activa mecanismos de cambio cuando deja de ser un requisito documental y se convierte en herramienta de orientación pedagógica. En los EE con mayor apropiación, el PFI permite identificar necesidades, ordenar prioridades, articular proyectos existentes y dar coherencia a acciones que antes podían encontrarse dispersas. También facilita la conversación entre directivos, docentes, tutores y ETC, pues ofrece un referente compartido para discutir avances, dificultades y ajustes.

El informe destaca que el PFI es el componente diferenciador del PTAFI 3.0 respecto de versiones anteriores del programa y que su operacionalización se refleja en efectos positivos sobre la articulación ETC–EE reportados por directivos y docentes. Esta evidencia sugiere que el PFI no solo cumple una función pedagógica, sino también una función de gobernanza: ayuda a alinear niveles institucionales y territoriales en torno a la formación integral.

No obstante, el mecanismo presenta riesgos de formalización. Cuando el PFI se elabora sin participación amplia, sin lectura contextual rigurosa o sin seguimiento institucional, puede convertirse en un documento de cumplimiento. En esos casos, su efecto sobre las prácticas pedagógicas se debilita. Por ello, la evaluación cualitativa subraya que el PFI tiene impacto cuando logra insertarse en la vida institucional y orientar decisiones concretas sobre tiempos, espacios, proyectos, responsabilidades y evaluación.

10.3.4. Resignificación del tiempo escolar y Centros de Interés

Esta lectura cualitativa converge con el hallazgo cuantitativo del AIPW/DR sobre docentes (subsección 9.2), que estima un efecto positivo y significativo del programa en siete de ocho dimensiones de transformación institucional, pero un efecto nulo sobre la dimensión Prácticas Pedagógicas individuales (+0,69 pp; $p > 0,10$). La formalización débil del PFI emerge así como uno de los mecanismos que explican el cuello de botella entre la transformación sistémica observada y la transformación didáctica de aula no observada.

El cuarto mecanismo es la resignificación del tiempo escolar mediante Centros de Interés y proyectos pedagógicos. Este mecanismo se activa cuando el tiempo escolar deja de organizarse únicamente alrededor de rutinas tradicionales de clase y se abre a experiencias formativas integrales vinculadas con arte, cultura, deporte, ciencia, tecnología, pensamiento histórico, lectura, oralidad, educación ciudadana, acción climática y saberes territoriales.

Los Centros de Interés tienen potencial transformador porque permiten reconocer intereses estudiantiles, diversificar mediaciones pedagógicas y conectar aprendizajes con experiencias significativas. En contextos donde el currículo se vive de manera rígida o fragmentada, estos espacios pueden ampliar las oportunidades de participación y aprendizaje. En particular, pueden favorecer la motivación estudiantil, la permanencia, el sentido de pertenencia y el vínculo escuela–comunidad.

Sin embargo, su impacto depende de la articulación curricular. Cuando los Centros de Interés se integran al PFI y a la planeación docente, pueden convertirse en escenarios de aprendizaje interdisciplinar. Cuando se implementan como actividades extracurriculares desconectadas de los propósitos formativos, su efecto se limita a la motivación temporal y no necesariamente incide en aprendizajes fundamentales.

La ruta del PTAFI 3.0 plantea precisamente la formación integral, los Centros de Interés y la resignificación del tiempo escolar como componentes centrales del nuevo enfoque programático, en articulación con lectura de contexto, equipos dinamizadores y educación CRESE. Desde la evaluación cualitativa, este mecanismo aparece como una de las apuestas más innovadoras del programa, pero también como una de las más exigentes, porque requiere tiempo, capacidades docentes, coordinación institucional y condiciones materiales.

10.3.5. Educación CRESE y fortalecimiento socioemocional

El quinto mecanismo es la incorporación de la educación CRESE como eje de formación ciudadana, reconciliación, antirracismo, desarrollo socioemocional y

acción climática. Este mecanismo se activa cuando la escuela incorpora de manera intencional prácticas de convivencia, reconocimiento de la diversidad, cuidado, participación, resolución pacífica de conflictos y construcción de ciudadanía.

CRESE opera en dos niveles. En el primero, introduce contenidos y orientaciones transversales que amplían la idea de aprendizaje más allá de lo cognitivo. En el segundo, modifica el clima relacional de la escuela al promover formas más dialogantes, inclusivas y sensibles a las trayectorias de los estudiantes. En contextos de vulnerabilidad, conflicto, ruralidad o exclusión histórica, este componente puede tener un valor estratégico para la permanencia y la convivencia.

El informe señala que CRESE aparece como mediador en maduración: los docentes reportan mejoras en la dimensión socioemocional de su práctica, aunque estas todavía no se reflejan plenamente en mediciones estandarizadas individuales. Esta lectura es consistente con una explicación cualitativa de proceso: las transformaciones socioemocionales suelen requerir continuidad, apropiación institucional y tiempo de sedimentación. No basta con introducir contenidos CRESE; se requiere que estos se traduzcan en relaciones pedagógicas, prácticas de aula, gestión de conflictos y cultura escolar.

Este mecanismo también permite explicar por qué algunos efectos pueden observarse antes en docentes y directivos que en estudiantes. La escuela puede estar transformando sus discursos, sus formas de acompañamiento y su sensibilidad frente al bienestar, pero los efectos en habilidades socioemocionales medibles pueden requerir ciclos más largos de exposición y evaluación.

10.3.6. Articulación territorial ETC–EE

El sexto mecanismo es la articulación entre las **Entidades Territoriales Certificadas** y los establecimientos educativos. Este mecanismo se activa cuando el programa no opera únicamente como intervención nacional, sino como proceso territorialmente acompañado, con capacidad de coordinación, seguimiento, apoyo y adaptación a las condiciones locales.

La articulación ETC–EE es relevante porque muchos de los factores que condicionan la implementación exceden la autonomía del EE: disponibilidad de docentes, transporte, conectividad, dotación, acompañamiento territorial, gestión de sedes rurales, formación complementaria y articulación intersectorial. Cuando la ETC participa activamente, el programa cuenta con mayores posibilidades de sostenibilidad. Cuando la relación es débil, el EE puede quedar solo frente a demandas que no puede resolver internamente.

Los resultados cuantitativos reportan efectos especialmente altos en la dimensión de articulación ETC–EE, particularmente en directivos. La lectura

cualitativa interpreta este hallazgo como señal de que el PTAFI 3.0 ha logrado fortalecer, al menos en parte, los canales de coordinación entre niveles de gestión. No obstante, esta articulación no es homogénea: depende de la capacidad técnica de la ETC, de la claridad de roles, de la continuidad de equipos y de la disponibilidad de recursos para acompañar el territorio.

10.4. Condiciones habilitantes y restrictivas

La activación de los mecanismos de cambio del PTAFI 3.0 depende de un conjunto de condiciones habilitantes y restrictivas. Estas condiciones no son externas al impacto, sino constitutivas de su posibilidad. En otras palabras, el programa no impacta de la misma manera en todos los EE porque no todos cuentan con las mismas capacidades institucionales, condiciones territoriales, trayectorias pedagógicas ni niveles de apropiación.

10.4.1. Condiciones habilitantes

La primera condición habilitante es la existencia de liderazgo directivo comprometido con la formación integral. Cuando el equipo directivo comprende el PTAFI 3.0 como una estrategia de transformación pedagógica y no como una obligación administrativa, se generan condiciones para articular el programa al PEI, al PMI, al PFI, al consejo académico y a los procesos de planeación institucional. En estos contextos, el tutor encuentra mayor legitimidad, los docentes participan con mayor disposición y los Centros de Interés pueden integrarse a la vida escolar.

La segunda condición habilitante es la estabilidad y disposición del cuerpo docente. La continuidad de los docentes permite que el acompañamiento tutorial tenga acumulación pedagógica. Si los docentes permanecen en el EE, las sesiones de trabajo situado, las orientaciones del tutor y los procesos de planeación pueden consolidarse progresivamente. Además, la disposición docente es fundamental: el programa requiere reflexión crítica sobre la práctica, apertura al cambio, capacidad de trabajo colaborativo y reconocimiento del tutor como par pedagógico.

La tercera condición es la existencia de una cultura institucional colaborativa. Los EE con experiencias previas de comunidades de aprendizaje, proyectos pedagógicos colectivos, participación estudiantil o articulación con familias muestran mayor capacidad de apropiación. En ellos, el PTAFI 3.0 no inicia desde cero, sino que potencia procesos existentes. La intervención se vuelve más efectiva cuando encuentra una base organizativa sobre la cual apoyarse.

La cuarta condición habilitante es la lectura de contexto. El programa adquiere pertinencia cuando sus acciones se diseñan desde las necesidades, intereses y condiciones reales de la comunidad educativa. La lectura de contexto permite que el PFI, los Centros de Interés y CRESE respondan a problemas

concretos: convivencia, permanencia, baja motivación, brechas lectoras, identidad cultural, cuidado ambiental o participación familiar. Sin lectura de contexto, la implementación corre el riesgo de reproducir formatos homogéneos poco sensibles al territorio.

La quinta condición es la articulación efectiva con la ETC. Allí donde la Secretaría de Educación acompaña, orienta, resuelve dificultades y alinea recursos, el programa tiene mayor capacidad de institucionalización. La ETC puede facilitar procesos de formación, apoyar la gestión de tutores, resolver asuntos logísticos, articular actores intersectoriales y promover sostenibilidad.

La sexta condición habilitante es la continuidad del acompañamiento. La intensidad y regularidad de la presencia tutorial resultan determinantes. Un acompañamiento episódico puede sensibilizar, pero difícilmente transforma prácticas. Un acompañamiento continuo puede generar confianza, seguimiento, retroalimentación y acumulación de aprendizajes institucionales.

La séptima condición es la disponibilidad de evidencias para la toma de decisiones. Los EE que usan resultados de evaluación, observación, asistencia, participación y desempeño estudiantil pueden orientar mejor sus acciones. En estos casos, el PTAFI 3.0 se conecta con una cultura de mejoramiento basada en información, no solo en percepciones generales.

10.4.2. Condiciones restrictivas

La primera condición restrictiva es la vulnerabilidad estructural de los territorios focalizados. El programa atiende EE con mayores desventajas acumuladas, especialmente en zonas rurales y municipios PDET. Esta focalización es coherente con un mandato de equidad, pero implica que los EE intervenidos parten de condiciones más difíciles que los controles. El informe documenta esta brecha estructural y muestra que los EE tratados concentran mayor proporción de municipios PDET, lo cual condiciona la interpretación de las brechas brutas de aprendizaje.

La segunda condición restrictiva es la alta rotación docente y directiva. Cuando los equipos cambian frecuentemente, los procesos de formación se interrumpen, se pierde memoria institucional y el tutor debe reiniciar ciclos de acompañamiento. Esta situación limita la acumulación necesaria para transformar prácticas.

La tercera condición restrictiva es la sobrecarga administrativa. En algunos EE, los actores perciben múltiples demandas simultáneas: diligenciamiento de formatos, atención a programas paralelos, reportes institucionales, exigencias curriculares, gestión de convivencia y resolución de problemas cotidianos. Esta

sobrecarga puede reducir el tiempo disponible para la reflexión pedagógica profunda.

La cuarta condición restrictiva es la débil articulación entre componentes del programa. Cuando el PFI, los Centros de Interés, CRESE y el acompañamiento tutorial se trabajan como líneas separadas, la intervención pierde coherencia. La formación integral requiere integración; si cada componente se implementa de manera aislada, su impacto se fragmenta.

La quinta condición restrictiva es la falta de condiciones materiales y logísticas. En zonas rurales dispersas, los problemas de transporte, conectividad, infraestructura, disponibilidad de espacios, dotación y tiempos de encuentro pueden limitar la implementación. Estos factores no anulan el potencial del programa, pero condicionan la intensidad real del tratamiento.

La sexta condición restrictiva es la persistencia de prácticas pedagógicas tradicionales. Aunque los docentes valoren el acompañamiento y reconozcan la importancia de la formación integral, la transformación de la práctica de aula es un proceso lento. Cambiar la planeación, la evaluación, la interacción, el uso del tiempo y la mediación didáctica requiere acompañamiento prolongado, retroalimentación situada y condiciones institucionales de apoyo.

La séptima condición restrictiva es la débil participación familiar en algunos contextos. La formación integral requiere corresponsabilidad, pero las familias enfrentan condiciones laborales, económicas, culturales o territoriales que pueden limitar su involucramiento. Esto afecta especialmente los componentes de permanencia, clima escolar y acompañamiento socioemocional.

En conjunto, estas condiciones permiten comprender por qué el programa muestra efectos importantes en dimensiones institucionales e intermedias, pero efectos más limitados o heterogéneos en aprendizajes estandarizados. El PTAFI 3.0 parece activar una cadena de transformación institucional, pero esta cadena encuentra cuellos de botella cuando debe traducirse en cambios didácticos sostenidos y resultados académicos individuales.

10.5. Casos contrastados: tratamiento/control o alta/baja intensidad

El análisis de casos contrastados permite profundizar en las diferencias cualitativas entre EE intervenidos y no intervenidos, así como entre EE con alta y baja intensidad de implementación del PTAFI 3.0. Esta estrategia no busca generalizar estadísticamente, sino construir explicaciones comparativas sobre las condiciones que favorecen o restringen el impacto.

En una evaluación basada en teoría, los casos contrastados cumplen tres funciones. Primero, permiten observar cómo opera la intervención en contextos

reales. Segundo, ayudan a identificar mecanismos causales que no siempre son visibles en los indicadores agregados. Tercero, permiten interpretar efectos heterogéneos, especialmente cuando EE con características similares muestran resultados distintos.

Para este capítulo, los casos pueden organizarse en cuatro integrales analíticos.

10.5.1. Caso tipo A: establecimiento educativo tratado con alta intensidad de implementación

Este tipo corresponde a EE intervenidos donde el PTAFI 3.0 ha logrado integrarse a la gestión institucional. Se caracteriza por liderazgo directivo activo, tutor reconocido por la comunidad educativa, docentes vinculados a procesos de planeación, PFI construido o actualizado de manera participativa, Centros de Interés en funcionamiento y articulación con la ETC.

En estos casos, el programa no se percibe como una visita externa, sino como parte de la vida institucional. El tutor acompaña, orienta y dinamiza; el directivo respalda; los docentes participan; los estudiantes reconocen nuevas experiencias formativas; y la comunidad empieza a identificar cambios en el clima escolar o en la motivación. El mecanismo predominante es la institucionalización del programa.

Los resultados esperados en este tipo de caso se observan primero en dimensiones intermedias: liderazgo, participación institucional, articulación ETC–EE, innovación pedagógica y clima escolar. Los efectos en aprendizajes pueden no aparecer de inmediato, pero existe mayor probabilidad de que la cadena causal avance hacia transformaciones de aula.

10.5.2. Caso tipo B: establecimiento educativo tratado con baja intensidad de implementación

Este tipo corresponde a EE formalmente intervenidos, pero con implementación débil, discontinua o fragmentada. Puede tratarse de sedes con visitas irregulares, baja disponibilidad de tiempo, débil liderazgo directivo, alta rotación docente, escasa apropiación del PFI o Centros de Interés desarrollados como actividades aisladas.

En estos casos, el programa llega, pero no logra convertirse en práctica institucional. Los actores pueden conocer el discurso de formación integral, pero no necesariamente traducirlo en decisiones curriculares o didácticas. El tutor puede ser valorado, pero su incidencia queda limitada por condiciones organizativas o logísticas. El mecanismo predominante es la sensibilización sin consolidación.

Los resultados esperados son parciales. Puede haber valoración positiva del programa y algunos cambios puntuales, pero no transformación sostenida. En términos cuantitativos, estos casos pueden contribuir a efectos promedio moderados o no significativos, especialmente en aprendizajes.

10.5.3. Caso tipo C: establecimiento educativo de control con capacidades institucionales previas

Este tipo corresponde a EE no intervenidos que, pese a no recibir el PTAFI 3.0 durante el periodo evaluado, cuentan con liderazgo pedagógico, cultura de planeación, proyectos institucionales activos, participación familiar o experiencias previas de mejoramiento. Estos establecimientos pueden presentar resultados iguales o superiores en algunos indicadores, no por ausencia de efecto del programa, sino porque parten de mejores condiciones institucionales.

Este tipo de caso es fundamental para evitar interpretaciones simplistas. Si un establecimiento de control tiene mejores resultados de fluidez lectora, ello puede reflejar ventajas previas, menor vulnerabilidad territorial o mayor estabilidad institucional. Por eso, el análisis contrafactual cuantitativo ajusta covariables, y el análisis cualitativo permite reconocer las diferencias de contexto que explican los desempeños.

En estas condiciones, el mecanismo predominante no proviene del PTAFI 3.0, sino de capacidades institucionales endógenas. La comparación con EE tratados vulnerables permite comprender la importancia de interpretar el impacto desde una perspectiva de equidad y no solo de diferencia bruta.

10.5.4. Caso tipo D: establecimiento educativo de control con baja capacidad institucional

Este tipo corresponde a EE no intervenidos que presentan debilidades en liderazgo, planeación, innovación, articulación territorial o clima escolar. En ellos, la ausencia del programa puede asociarse con menores oportunidades de acompañamiento pedagógico situado y menor acceso a dispositivos de formación integral.

La comparación entre este tipo de control y EE tratados de características similares permite identificar el valor agregado del PTAFI 3.0. Si los EE tratados muestran mayor liderazgo distribuido, articulación institucional o participación, ello sugiere que el programa activa mecanismos que no emergen espontáneamente en contextos de baja capacidad.

Este tipo de contraste es especialmente relevante para interpretar los efectos positivos en actores educativos. Aunque la mejora en aprendizajes estandarizados

pueda ser limitada en el corto plazo, el programa puede estar generando condiciones institucionales que los EE de control vulnerables no tienen.

10.5.5. Lectura comparativa de los casos

La comparación entre casos permite formular una conclusión sustantiva: la diferencia clave no está únicamente en recibir o no recibir el programa, sino en la intensidad y calidad de su apropiación. El tratamiento nominal no equivale necesariamente al tratamiento efectivo. Un EE puede estar focalizado, pero recibir una dosis baja o fragmentada de implementación. Otro puede no estar focalizado, pero contar con capacidades previas que le permiten obtener buenos resultados.

La diferenciación entre niveles de intensidad y apropiación se sustenta en dos tipos de evidencia recogida en el componente cualitativo. Primero, los registros etnográficos de campo documentan diferencias observables en la frecuencia de visitas tutoriales (rango: una visita semanal en EE de alta intensidad frente a una visita mensual o quincenal en EE de baja intensidad), en el grado de integración del Plan de Formación Integral en la planeación institucional (PFI operativo con seguimiento periódico vs. PFI archivado), y en la activación efectiva de Centros de Interés (presencia regular en horario escolar vs. activaciones puntuales). Segundo, las entrevistas con directivos y docentes evidencian variaciones en la apropiación discursiva del programa: en EE de alta apropiación, los actores describen el programa en primera persona y como parte de su identidad institucional; en EE de baja apropiación, lo describen como un proyecto externo del MEN al que se rinden informes. Esta evidencia, sistematizada en la matriz CMO del Anexo E, sustenta la distinción entre tratamiento nominal y tratamiento efectivo.

Por ello, en la tabla continua la evaluación cualitativa propone distinguir entre cuatro niveles de intensidad:

Tabla 10-2

Niveles de la evaluación cualitativa

Nivel de intensidad	Caracterización	Resultado cualitativo probable
Alta intensidad	Tutor activo, liderazgo directivo, PFI apropiado, Centros de Interés articulados, ETC presente	Transformación institucional y condiciones para cambio pedagógico
Intensidad media	Acciones implementadas, pero con integración parcial	Cambios intermedios, apropiación desigual
Baja intensidad	Cumplimiento formal, visitas discontinuas, débil apropiación	Sensibilización sin transformación sostenida

Sin tratamiento efectivo	Ausencia del programa o implementación marginal	Dependencia de capacidades previas del EE.
---------------------------------	---	--

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

Esta lectura es fundamental para explicar los efectos heterogéneos. El impacto del PTAFI 3.0 no depende únicamente de la asignación al grupo tratado, sino de la calidad de la mediación tutorial, la apropiación institucional y la capacidad del contexto para convertir recursos programáticos en prácticas sostenibles.

10.6. Explicación cualitativa de efectos, no efectos y efectos heterogéneos

La articulación entre los tipos de efecto identificados cuantitativamente y los mecanismos cualitativos puede sintetizarse así. (i) Efectos positivos: se observan en eficiencia interna (PDO1, IRI5, IRI17) y en siete de ocho dimensiones docentes del AIPW/DR (capítulo 9). Cualitativamente se explican por el mecanismo del acompañamiento tutorial sostenido y por la activación del Plan de Formación Integral cuando logra insertarse en la vida institucional. (ii) No efectos: el ATE EGRA (-3,66 pp; $p = 0,011$ en Fase I y -4,31 pp; $p = 0,368$ en Fase II) y el ATE EGMA no significativo se explican cualitativamente por el cuello de botella entre transformación institucional y didáctica de aula, y por la dosificación insuficiente de modelado tutorial en lectura inicial. (iii) Efectos heterogéneos: la magnitud diferencial del CATE en rural (-11,00 pp) y PDET (-40,13 pp) se explica por la articulación territorial débil entre Entidades Territoriales Certificadas y establecimientos educativos, por la rotación de tutores y por la discontinuidad de operativos durante la pandemia. Esta desagregación por tipo de efecto se presenta de forma sistemática en la matriz CMO consolidada (Anexo E) y orienta las recomendaciones del capítulo 14.

La evaluación cualitativa de impacto permite interpretar tres tipos de resultados: efectos positivos, no efectos y efectos heterogéneos. Esta distinción es central para evitar lecturas reduccionistas de la evidencia. En programas educativos complejos, un efecto no significativo en una prueba estandarizada no implica necesariamente ausencia de transformación; puede indicar que la cadena causal está activa en niveles intermedios, pero aún no ha alcanzado el resultado final esperado. Del mismo modo, un efecto positivo en liderazgo o clima escolar no garantiza automáticamente mejora inmediata en fluidez lectora, aunque puede constituir una condición necesaria para alcanzarla en el mediano plazo.

10.6.1. Explicación cualitativa de los efectos positivos

Los efectos positivos del PTAFI 3.0 se explican principalmente por la activación de mecanismos institucionales. La evidencia cuantitativa reportada en el informe muestra efectos significativos en varias dimensiones de directivos, docentes, estudiantes y acudientes, con particular fuerza en articulación ETC–EE,

liderazgo directivo, participación institucional, participación comunitaria e innovación pedagógica.

Desde el análisis cualitativo, estos efectos pueden explicarse por cuatro procesos. Primero, la presencia del tutor introduce una mediación pedagógica que ayuda a ordenar conversaciones institucionales sobre enseñanza, formación integral y mejoramiento. Segundo, el PFI ofrece un dispositivo de planificación que permite articular acciones dispersas y orientar decisiones. Tercero, la formación integral amplía el sentido de la escuela y legitima dimensiones socioemocionales, ciudadanas, culturales y territoriales. Cuarto, la articulación con ETC y equipos dinamizadores fortalece la gobernanza del programa.

Estos efectos son coherentes con la naturaleza del PTAFI 3.0. El programa no actúa únicamente sobre estudiantes individuales, sino sobre estructuras institucionales de mediación pedagógica. Por ello, sus primeros impactos aparecen en actores adultos y en condiciones organizativas. Esto no debe ser interpretado como efecto menor; al contrario, indica que la intervención está modificando los entornos que hacen posible la enseñanza y el aprendizaje.

La mejora en liderazgo, articulación y participación puede entenderse como impacto de primer orden. La transformación de prácticas docentes sería un impacto de segundo orden. La mejora en aprendizajes estandarizados sería un impacto de tercer orden, más dependiente de tiempo, continuidad, intensidad didáctica y condiciones estructurales.

10.6.2. Explicación cualitativa de los no efectos

Los no efectos, especialmente en aprendizajes estandarizados o en algunas dimensiones socioemocionales medidas individualmente, pueden explicarse por la presencia de cuellos de botella (entendidos como puntos críticos donde la transformación institucional no logra traducirse en transformación didáctica observable en el aula, interrumpiendo la cadena que conecta liderazgo, planeación y prácticas pedagógicas) en la cadena causal. El principal cuello de botella identificado es la traducción de cambios institucionales en transformación didáctica sostenida.

El programa puede fortalecer la planeación institucional, el liderazgo y la participación, pero si estos procesos no llegan al aula de manera sistemática, su efecto sobre fluidez lectora será limitado. La fluidez lectora requiere práctica intensiva, retroalimentación frecuente, seguimiento individual, materiales adecuados, tiempo pedagógico efectivo y estrategias específicas de alfabetización. La formación integral crea condiciones favorables, pero no sustituye la necesidad de intervención didáctica explícita en lectura.

Otro factor que explica los no efectos es el horizonte temporal. Las transformaciones pedagógicas requieren ciclos prolongados. Un programa puede mostrar efectos rápidos en percepción, coordinación y clima, pero los aprendizajes acumulados demandan mayor tiempo de exposición. En EE con rezagos estructurales, el primer logro puede ser estabilizar trayectorias, mejorar asistencia, reducir reprobación o fortalecer condiciones institucionales, antes que elevar puntajes de lectura.

También debe considerarse la focalización del programa. El PTAFI 3.0 atiende EE con mayores desventajas. Por ello, una brecha negativa persistente en fluidez lectora no necesariamente indica efecto adverso del programa; puede expresar la profundidad de las condiciones de partida. El propio informe interpreta la brecha de EGRA desde una perspectiva de equidad: el programa atiende EE con mayor vulnerabilidad y, aunque no cierra de inmediato la brecha individual de aprendizajes, sí muestra señales de convergencia en eficiencia interna y fortalecimiento institucional.

En términos cualitativos, esto puede describirse como un efecto de contención y organización. El programa ayuda a sostener, ordenar y fortalecer procesos en contextos adversos, aunque todavía no logre revertir plenamente las brechas académicas acumuladas. La ausencia de efecto inmediato en aprendizajes no invalida la intervención, pero sí señala la necesidad de profundizar su componente didáctico específico, especialmente en lectura, escritura, pensamiento matemático y evaluación formativa.

10.6.3. Explicación cualitativa de los efectos heterogéneos

Los efectos heterogéneos se explican por la interacción entre contexto, intensidad de implementación y mecanismos activados. En EE rurales, PDET o con mayor vulnerabilidad, el programa enfrenta condiciones más restrictivas, pero también puede tener mayor valor agregado institucional. Esto genera una paradoja evaluativa: los EE más necesitados pueden mostrar menores resultados académicos absolutos, pero mayor transformación relativa en procesos institucionales y trayectorias escolares.

La heterogeneidad también se relaciona con la capacidad de apropiación. Dos EE tratados pueden recibir orientaciones similares, pero producir resultados distintos si uno cuenta con liderazgo directivo, docentes estables y cultura colaborativa, mientras otro enfrenta rotación, baja participación y dificultades logísticas. En el primer caso, el mecanismo se activa; en el segundo, se debilita o queda incompleto.

Asimismo, la heterogeneidad puede explicarse por diferencias en la intensidad tutorial. La continuidad del tutor, la calidad de la relación pedagógica, la

frecuencia de encuentros y la capacidad de adaptación al contexto inciden en el nivel de apropiación. Allí donde el tutor logra construir confianza y legitimidad, el programa se vuelve más efectivo. Allí donde su presencia es intermitente o se percibe como externa, su incidencia disminuye.

Los efectos heterogéneos también dependen del grado de integración entre PFI, CRESE y Centros de Interés. En EE donde estos componentes se articulan, la formación integral se convierte en estrategia institucional. Donde se implementan de manera separada, los efectos se dispersan.

Esta integración opera a través de tres condiciones observadas en los casos de mayor apropiación. Primero, el PFI funciona como instrumento rector que articula las decisiones sobre tiempos, espacios y responsabilidades; segundo, el CRESE actúa como mecanismo de cohesión comunitaria que da legitimidad social al proyecto educativo y vincula a las familias y a las organizaciones territoriales; tercero, los Centros de Interés materializan la resignificación del tiempo escolar al traducir las orientaciones del PFI en experiencias pedagógicas concretas, frecuentemente vinculadas a saberes territoriales, arte, deporte y ciencia. En los establecimientos donde estos tres componentes se implementan de forma desarticulada, cada uno opera como actividad aislada y los efectos se diluyen. En los establecimientos donde se articulan, la formación integral se consolida como estrategia institucional sostenible.

10.6.4. Síntesis interpretativa

La explicación cualitativa de impacto permite formular una lectura integrada del PTAFI 3.0: El programa muestra mayor capacidad de impacto en el plano institucional que en el plano estrictamente didáctico individual. Ha logrado activar mecanismos de liderazgo distribuido, articulación territorial, apropiación inicial del PFI, fortalecimiento del clima escolar y apertura hacia la formación integral. Sin embargo, la transformación de las prácticas docentes de aula y su traducción en mejoras estandarizadas de aprendizaje constituyen todavía el tramo más débil de la cadena causal.

Esta conclusión no debe leerse como fracaso del programa, sino como identificación precisa de su momento de maduración. El PTAFI 3.0 está generando condiciones institucionales necesarias para la mejora educativa, pero requiere fortalecer la especificidad didáctica, la continuidad del acompañamiento y el uso pedagógico de evidencias para que dichas condiciones se traduzcan en aprendizajes fundamentales.

En términos de teoría de cambio, la cadena causal puede representarse así:

Tutoría situada + liderazgo directivo + PFI + CRESE + Centros de Interés + articulación ETC–EE → fortalecimiento institucional y clima escolar → transformación gradual de prácticas docentes → mejora esperada de aprendizajes y trayectorias escolares.

El hallazgo cualitativo central es que los primeros eslabones de la cadena están activos, especialmente liderazgo, articulación y formación integral; los eslabones intermedios muestran activación parcial, particularmente prácticas docentes y apropiación curricular; y el eslabón final —aprendizajes estandarizados— requiere mayor tiempo, intensidad y focalización didáctica.

Por tanto, la evaluación cualitativa de impacto basada en teoría permite afirmar que el PTAFI 3.0 opera como un dispositivo de equidad institucional y pedagógica. Su principal contribución en esta fase no es aún la mejora inmediata y homogénea de todos los indicadores de aprendizaje, sino la creación de condiciones para que dicha mejora sea posible: coordinación institucional, liderazgo distribuido, acompañamiento situado, formación integral, resignificación del tiempo escolar y mayor sensibilidad frente a las realidades territoriales.

En consecuencia, la recomendación derivada de este capítulo es fortalecer la transición entre transformación institucional y transformación didáctica. Para ello, se requiere profundizar el acompañamiento en aula, orientar los Centros de Interés hacia aprendizajes fundamentales, vincular CRESE con prácticas pedagógicas evaluables, consolidar el PFI como instrumento vivo de gestión curricular y asegurar que la articulación ETC–EE se traduzca en apoyos concretos para la enseñanza. Solo así el programa podrá convertir sus efectos institucionales ya observables en impactos académicos sostenidos, especialmente en los EE rurales, PDET y de mayor vulnerabilidad educativa.

11. Integración de hallazgos de impacto

Esta sección consolida la lectura integrada de los hallazgos cuantitativos (capítulos 7-9) y cualitativos (capítulo 10) y constituye el insumo principal para la formulación de las recomendaciones de política del capítulo 14. Por su carácter sintético y articulador, sus contenidos centrales se trasladan al resumen ejecutivo del informe, donde se presentan en clave de cuatro patrones de convergencia y sus implicaciones directas de política pública.

Tres ejes vertebran la lectura integrada del impacto del PTAFI 3.0. El primero es la convergencia entre los efectos detectados por la inferencia causal AIPW/DR y la explicación narrativa de los actores en el operativo cualitativo, que se desarrolla en la subsección 11.1. El segundo es la identificación de los factores mediadores que conectan la presencia del tutor con los resultados sobre dimensiones intermedias y, eventualmente, sobre aprendizajes estandarizados, que se desarrolla en la subsección 11.2. El tercero es la pregunta de evaluación de impacto basada en teoría, que indaga qué funcionó, para quiénes, en qué contextos y por qué, y que se desarrolla en la subsección 11.3 articulando las cinco hipótesis cualitativas de lectura del informe Fase I con los hallazgos cuantitativos del Informe (15).

Este capítulo articula los hallazgos cuantitativos del modelo AIPW/DR, los descriptivos longitudinales de las Tasas de Eficiencia Interna y la replicación de los indicadores PROMISE con la evidencia cualitativa de procesos generada por el Equipo Cualitativo en grupos focales, entrevistas semiestructuradas y observación etnográfica de aulas. La integración sigue el enfoque de complementariedad analítica de Greene (2007) y la triangulación metodológica de Pawson y Tilley (1997), conjugando las preguntas que cada método responde de manera privilegiada: el componente cuantitativo identifica QUÉ efectos ocurren y CUÁNTO miden, el componente cualitativo explica CÓMO operan los mecanismos causales y POR QUÉ se observan patrones diferenciales según contexto.

11.1. Convergencia entre inferencia causal y explicación cualitativa

El primer patrón de convergencia es la posición estructural desfavorable de los EE tratados. La evidencia cuantitativa documenta que en 2017 los tratados aprobaban -1,79 puntos porcentuales menos que los controles, brecha que se amplió a -2,27 pp en 2022 por efecto pandémico y se redujo a -0,82 pp en 2025 (subsección 8.1.1). En municipios PDET la brecha se invirtió a +0,66 pp en 2025. La evidencia cualitativa explica el origen de esta posición: los directivos y tutores describen EE focalizados como sedes con menor capital humano docente disponible, mayor rotación, infraestructura precaria y población estudiantil con

cargas de vulnerabilidad multidimensional acumulada. La convergencia es directa: el programa atiende a quienes más necesitan acompañamiento, con una línea de base más baja que genera apariencia de bajo desempeño relativo cuando se compara con cualquier referente.

Esta convergencia cuanti-cualitativa sobre la posición estructural desfavorable de los EE tratados constituye uno de los hallazgos centrales de la evaluación y se incorpora de forma explícita en el resumen ejecutivo del informe como elemento de lectura obligada para la interpretación correcta de los efectos estimados.

El segundo patrón es la transformación sistémica institucional sin transformación didáctica individual. El AIPW/DR sobre docentes (subsección 9.2) identifica efectos positivos en 7 de 8 dimensiones, con magnitudes mayores en participación institucional (+9,54 pp), liderazgo directivo (+8,60 pp) y participación comunitaria (+7,17 pp), pero efecto nulo en prácticas pedagógicas individuales (+0,69 pp, no significativo). La evidencia cualitativa lo explica: los docentes describen al tutor como una figura que reorganiza la coordinación pedagógica, introduce reuniones de planeación conjunta, sostiene la cultura institucional de evaluación formativa y articula la sede con la Entidad Territorial Certificada, pero raras veces ingresa al aula a modelar didácticamente y no impone metodologías específicas. La convergencia sustenta la lectura central del informe Fase I de "contención sin transformación": el programa estabiliza el entorno y mantiene la matrícula, pero no produce todavía las transformaciones didácticas necesarias para mover los aprendizajes estandarizados.

Este patrón se retoma en el capítulo 14 (Recomendación Primera) como fundamento del refuerzo del enfoque territorial diferencial, especialmente en zonas rurales y municipios PDET, donde la posición estructural inicial demanda un escalonamiento progresivo y una focalización sostenida del acompañamiento tutorial.

El tercer patrón es el efecto pandémico diferencial. El análisis cuantitativo documenta el máximo de la brecha T-C en 2022 (-2,27 pp en aprobación total, +1,13 pp en deserción) y la recuperación más lenta en EE rurales y PDET. La evidencia cualitativa explica que en zonas rurales y PDET la conectividad limitada, la migración económica de hogares y la discontinuidad de tutores durante 2020-2021 amplificaron el shock pandémico, mientras que en zonas urbanas el retorno a presencialidad fue más rápido. La convergencia indica que la recuperación 2023-2025 es atribuible al menos parcialmente al programa, no solo al efecto natural de retorno presencial: los tratados PDET son los que más se recuperan, lo cual no sería esperable si el efecto fuera solo macroeducativo.

La evidencia cualitativa que sustenta la contribución parcial del programa a la recuperación post-pandémica proviene de tres fuentes triangulares. Primero, las entrevistas a tutores documentan que en zonas PDET y rurales la presencia tutorial continuada durante 2023-2025 actuó como mecanismo de estabilización institucional frente a la desarticulación previa. Segundo, las matrices CMO identifican que los EE tratados con menor rotación de tutor logran reactivar más rápidamente reuniones de planeación conjunta y evaluación formativa, prácticas suspendidas durante la pandemia. Tercero, los datos cuantitativos muestran una recuperación más marcada de la brecha T-C en aprobación y deserción entre 2022 y 2025 en PDET (de -1,79 pp a +0,66 pp) que en otras zonas, magnitud que excede lo esperable por efecto natural de retorno presencial. Esta convergencia direcciona la formulación de la Recomendación Tercera (capítulo 14) sobre sostenibilidad del acompañamiento tutorial en zonas post-pandémicas.

El cuarto patrón es la paradoja percepción-resultado en aprendizajes. La estimación AIPW/DR de la Fase I encuentra un efecto promedio negativo significativo sobre fluidez lectora EGRA (-3,66 pp, $p=0,011$) y nulo sobre EGMA (-0,06 pp, ns). En cambio, los actores adultos (directivos, docentes) y los estudiantes auto-reportan mejoras significativas en aprendizajes percibidos (DD +7,66 pp, DO +3,60 pp, ES +2,52 pp). La evidencia cualitativa identifica el mecanismo del desfase: los actores reportan mejoras en hábitos de estudio, asistencia, participación en clase y motivación, pero estas mejoras no se traducen aun en logros estandarizados de fluidez lectora porque la transformación didáctica del aula es lenta y los docentes mantienen rutinas previas. La convergencia direcciona la recomendación de política hacia el fortalecimiento del componente de aprendizajes (capítulo 14, Recomendación Cuarta).

El quinto patrón es la homogeneidad operativa de los tutores frente a la heterogeneidad territorial. El análisis intra-tratamiento de tutores (subsección 9.3) identifica medias auto-reportadas entre 75 y 88 puntos en escala 0-100 sin diferencias estadísticamente significativas por zona, PDET o experiencia. La evidencia cualitativa lo explica: el cuerpo de tutores ha desarrollado una cultura profesional homogénea con protocolos compartidos, comunidades de práctica regionales y supervisión por microrregión, lo cual genera estándares uniformes de actuación que constituyen un activo del programa. La convergencia sustenta el argumento de que el tutor es el mecanismo causal central del PTAFI 3.0.

La convergencia entre la inferencia causal cuantitativa y la explicación cualitativa se manifiesta en cinco patrones que se sostienen sobre fuentes independientes: el panel oficial TEI 2017-2025 de 8.455 establecimientos educativos, la submuestra EGRA/EGMA Fase I de 22 EE y 776 niños, las cinco encuestas CN del operativo Fase II 2026 sobre 408 EE y 2.431 actores, y los

registros etnográficos de 9 rutas de visita. La Tabla 11-1 y la Figura 11-1 sintetizan la convergencia hallazgo por hallazgo.

Tabla 11-1.

Convergencia entre la inferencia causal cuantitativa y la explicación cualitativa, organizada por patrón y por mecanismo identificado.

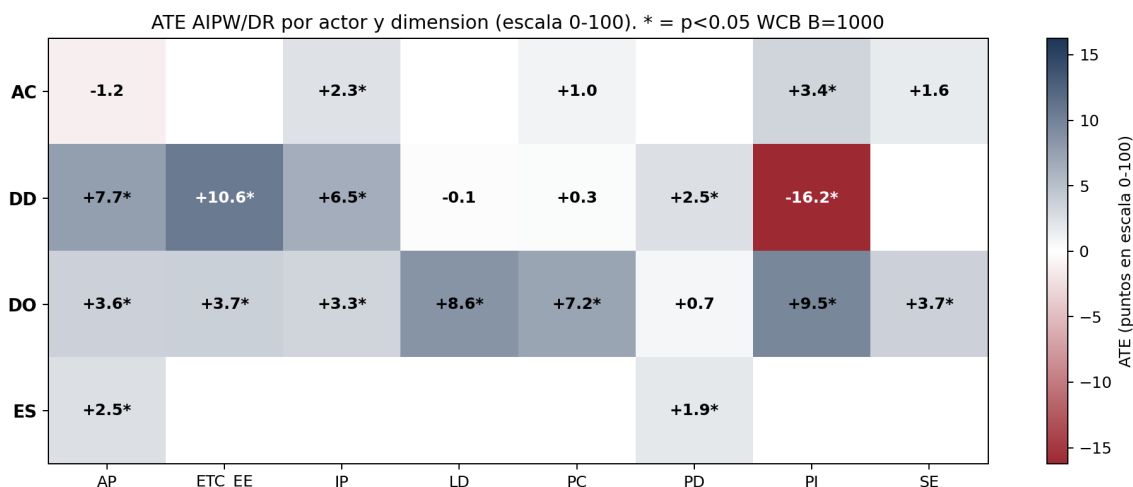
Patrón	Evidencia cuantitativa	Explicación cualitativa	Mecanismo identificado
1. Posición estructural desfavorable	TEI: brecha T-C aprobación 2017=-1,79, 2025=-0,82; PDET inversión +0,66	Sedes focalizadas con menor capital docente, mayor rotación, vulnerabilidad multidimensional	Focalización deliberada en territorios más necesitados
2. Transformación sistémica sin transformación didáctica	DO: AP +3,60, LD +8,60, PI +9,54, PC +7,17, ETC_EE +3,72; PD +0,69 ns	Tutor reorganiza coordinación pedagógica pero raras veces ingresa al aula	Coaching institucional con bajo dosaje de modelado didáctico individual
3. Efecto pandémico diferencial	Máximo brecha 2022, recuperación 2023-2025 mayor en PDET tratados	Conectividad limitada y discontinuidad de tutores 2020-2021 en zonas vulnerables	Reactivación pos-pandémica acelerada por presencia continuada del tutor
4. Paradoja percepción-resultado	EGRA Fase I -3,66 (sig); DD AP +7,66, DO AP +3,60, ES AP +2,52	Mejora hábitos, asistencia, motivación; rutinas didácticas mantenidas	Desfase temporal entre cambios de proceso y logros estandarizados
5. Homogeneidad operativa de tutores	TU intra-tratamiento medias 75-88, sin diferencias por zona/PDET/experiencia	Cultura profesional homogénea, protocolos compartidos, supervisión regional	Tutor como vector causal central con estándares uniformes territorialmente

Nota aclaratoria: los mecanismos listados en la columna anterior corresponden a la categorización cualitativa específica del análisis de casos del capítulo 10 (subsección 10.3), y deben distinguirse de los ocho mecanismos de cambio del modelo causal documentados en el capítulo 9 (subsección 9.1, mecanismos M1 a M8). Ambos sistemas son complementarios pero no equivalentes: los mecanismos del capítulo 9 corresponden a construcciones agregadas validadas con AIPW/DR, mientras que los del capítulo 10 corresponden a hallazgos emergentes de los casos cualitativos. La matriz de correspondencia entre ambos sistemas se presenta en el Anexo E (matriz CMO consolidada).

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Elaboración propia integrando AIPW/DR Fase I (cap. 7), TEI 2017-2025 (cap. 8), AIPW/DR actores (cap. 9) y registros cualitativos (cap. 10).

Figura 11-1.

Mapa Cross-actor del ATE estimado por dimensión y actor, base para la lectura integrada de convergencia. Las celdas con * corresponden a $p < 0,05$ WCB.



Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026.

11.2. Factores mediadores: liderazgo, prácticas, clima, PFI, CRESE.

La Tabla 11-2 sintetiza los cinco mediadores con su magnitud cuantitativa, su evidencia cualitativa de soporte y el nivel de activación observado. La conclusión del análisis mediador es que el programa opera mediante una cadena causal liderazgo distribuido a clima escolar y articulación ETC-EE a participación comunitaria a CRESE, en la que las prácticas docentes individuales y los aprendizajes estandarizados constituyen el eslabón final aun no plenamente alcanzado.

Competencias socioemocionales y ciudadanas (CRESE). El instrumento Quiero Ser Quiero Saber (QSQS) aplicado en grados 5° y 9° (subsección 8.5.4) reporta puntajes equivalentes en EE PTAFI 3.0 y nacionales en competencias ciudadanas y habilidades socioemocionales. La integración con el AIPW/DR de actores indica que las competencias socioemocionales operan como mediador parcial: los docentes identifican mejoras significativas en la dimensión socioemocional de su práctica (+3,66 pp), pero estas mejoras no se traducen aun en mediciones CRESE estandarizadas, lo que sugiere que el mediador está activo en el aula, pero requiere tiempo de maduración para reflejarse en los puntajes individuales de los estudiantes.

Plan de Formación Integral (PFI). El PFI es el componente diferenciador del PTAFI 3.0 respecto de las versiones anteriores del programa y constituye el mediador específico de la versión actual. La dimensión Articulación ETC-EE captura su operacionalización administrativa y arroja efectos positivos en los dos actores que la reportan: directivos (+10,57 pp) y docentes (+3,72 pp). La evidencia

cualitativa identifica al PFI como el dispositivo que ha permitido que el acompañamiento del tutor trascienda la primaria y se extienda a transición, secundaria y media; este mediador explica la coherencia metodológica de adoptar la especificación UDENAR (todos los niveles) en lugar de la PROMISE (solo P+S) en el análisis TEI (subsección 8.1.5).

Clima escolar. El clima escolar se aproxima por la dimensión Socioemocional medida en docentes (+3,66 pp, sig) y acudientes (+1,63 pp, ns), y por la dimensión Participación comunitaria medida en docentes (+7,17 pp, sig) y acudientes (+0,97 pp, ns). La triangulación sugiere que los actores escolares perciben un clima institucional más estable y más conectado con el entorno familiar, mientras que los acudientes detectan estos cambios con menor intensidad. La recuperación en deserciones (capítulo 8) y la disminución de la brecha T-C en deserciones a +0,54 pp en 2025 son consistentes con un mediador clima escolar activo: el programa retiene estudiantes que en su contrafactual habrían abandonado el sistema.

Prácticas docentes. La evidencia cuantitativa identifica un ATE no significativo en prácticas docentes auto-reportadas por los propios docentes (PD +0,69 pp, $p=0,390$), pero significativo en prácticas observadas por los estudiantes (ES PD +1,88 pp) y por los directivos (DD PD +2,46 pp). La discrepancia direccional entre auto-reporte docente y observación externa es un artefacto conocido en la literatura de evaluación educativa (OECD, 2018) y refleja sesgo de deseabilidad social en el primero. La lectura integrada concluye que el programa produce cambios incipientes en prácticas pedagógicas observables (planeación conjunta, evaluación formativa, retroalimentación), pero estos cambios todavía no han sido apropiados por los docentes como transformación identitaria de su rol. Este mediador es el de menor magnitud efectiva y representa el principal cuello de botella para el impacto sobre aprendizajes estandarizados.

Liderazgo directivo distribuido. El AIPW/DR sobre docentes identifica un ATE de +8,60 pp en liderazgo directivo, mientras que sobre directivos el mismo concepto se mide en sentido inverso por la dimensión Participación institucional (-16,25 pp en directivos, +9,54 pp en docentes). La interpretación conjunta es coherente con la teoría de liderazgo distribuido (Hattie, 2009): el directivo de un EE tratado delega funciones de moderación pedagógica al tutor y al cuerpo docente, lo que reduce su frecuencia de intervención directa en instancias institucionales, pero amplía el ejercicio del liderazgo a otros actores. Este mediador es el de magnitud más alta del modelo y constituye la primera ruta de impacto del programa: el tutor reorganiza la estructura de relaciones de autoridad pedagógica en el EE.

La identificación de factores mediadores responde a la pregunta de **CÓMO opera el efecto del programa**: que mecanismos conectan la presencia del tutor con los cambios observados en las dimensiones de actores y, eventualmente, con los resultados sobre aprendizajes y trayectorias escolares. La evaluación adopta un modelo de rutas de complejidad media (path analysis con variables observadas, conforme al Informe Metodológico, MEN-UDENAR, 2025) que organiza los mediadores en cinco grupos: liderazgo directivo distribuido, prácticas docentes, clima escolar, Plan de Formación Integral (PFI) y competencias socioemocionales (CRESE).

Tabla 11-2.

Cinco mediadores de la cadena causal del PTAFI 3.0: magnitud cuantitativa, evidencia cualitativa y nivel de activación.

Mediador	Magnitud cuantitativa (pp en escala 0-100)	Evidencia cualitativa	Nivel de activación
Liderazgo directivo distribuido	DO LD +8,60 sig; DD PI -16,25 sig (interpretación: distribución del liderazgo)	Tutor reorganiza coordinación pedagógica; consejos académicos rotan moderación	Alto: primer eslabón causal
Prácticas docentes	DO PD +0,69 ns; ES PD +1,88 sig; DD PD +2,46 sig (auto-reporte docente bajo)	Cambios incipientes en planeación conjunta y evaluación formativa; rutinas mantenidas	Bajo: cuello de botella
Clima escolar	DO SE +3,66 sig; DO PC +7,17 sig; AC SE +1,63 ns	Más estabilidad institucional; familias menos involucradas que docentes	Medio-alto
Plan de Formación Integral (PFI)	DD ETC_EE +10,57 sig; DO ETC_EE +3,72 sig	Componente diferenciador del PTAFI 3.0; cubre transición-secundaria-media	Alto: mediador específico
CRESE socioemocional	DO SE +3,66 sig; QSQS sin diferencia significativa con benchmark nacional	Mejora en aula reportada por docentes; no reflejada aun en mediciones individuales	Medio: mediador en maduración

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Integración de AIPW/DR cap. 9, QSQS cap. 8.5, hallazgos cualitativos cap. 10.

11.3. ¿Qué funcionó, para quiénes, en qué contextos y por qué?

La Tabla 11-3 sintetiza la respuesta a las cuatro preguntas en una matriz integrada. La lectura central del informe (15) puede formularse en una única oración: el PTAFI 3.0 cierra la brecha estructural de eficiencia interna en zonas vulnerables y transforma la organización institucional de los EE focalizados, pero todavía no logra producir transformaciones didácticas particulares que se traduzcan en mejoras sustantivas en logros estandarizados de aprendizaje.

La explicación causal integrada del por qué es la siguiente: el PTAFI 3.0 actúa como un dispositivo de transformación sistémica institucional mediado por un tutor con alto desempeño profesional homogéneo en el territorio. El tutor reorganiza la coordinación pedagógica del EE, redistribuye funciones de liderazgo entre directivo y cuerpo docente, sostiene una cultura institucional de evaluación formativa y conecta la sede con la Entidad Territorial Certificada vía PFI. Esta cadena causal produce efectos detectables en las dimensiones intermedias (LD, PI, PC, ETC_EE, SE) en plazos de uno a tres años, y en las tasas de eficiencia interna en plazos de tres a cinco años. Los efectos sobre aprendizajes estandarizados requieren plazos más largos porque dependen de la transformación didáctica individual de los docentes, que es el eslabón causal más resistente al cambio. La recomendación de política para acelerar este eslabón final se desarrolla en el capítulo 14 (Recomendación Cuarta).

¿En qué contextos? Tres contextos institucionales facilitan los efectos positivos del programa. Primero, los EE con directivos dispuestos a delegar funciones de coordinación pedagógica al tutor: el liderazgo directivo abierto al coaching es la condición habilitante más robusta. Segundo, las microrregiones con cuerpos de tutores de baja rotación: la continuidad temporal del mismo tutor en la misma sede es el predictor más fuerte de la profundidad de los cambios institucionales observados. Tercero, los municipios donde el PTAFI 3.0 se articula con políticas complementarias (Programa de Alimentación Escolar, Familias en Acción, Jornada Única): la concurrencia de programas amplifica los efectos del tutor sobre la retención y la participación comunitaria. Tres contextos restringen los efectos: rotación alta de docentes que diluye el efecto acumulado del coaching, discontinuidad de financiamiento entre vigencias que interrumpe el ciclo de acompañamiento, y conflicto territorial activo que dificulta la operación en algunas zonas D y E.

¿Para quiénes? El programa muestra heterogeneidad de impacto en cinco dimensiones de subgrupo. Por tipo de actor: el cuerpo docente concentra los efectos más grandes y consistentes (7/8 dimensiones positivas), mientras que estudiantes y acudientes detectan efectos modestos y parciales. Por zona geográfica: la convergencia de la brecha T-C es más acelerada en zonas rurales (-1,57 a -1,06 pp) que en zonas urbanas (-2,13 a -0,46 pp), y especialmente acelerada en municipios PDET donde se invierte a favor del programa. Por nivel educativo: la inclusión de transición y media en la cobertura del programa vía PFI explica la coherencia entre las series UDENAR (todos los niveles) y PROMISE (P+S) (subsección 8.1.5). Por exposición acumulada: aunque el análisis por cohorte se reportará al cierre del operativo (subsección 8.2), la evidencia BM/DNP 2022 sugiere efectos crecientes con mayor antigüedad en el programa. Por capacidad

institucional inicial: los EE con mejor history TEI 2017-2024 son los que más se benefician del componente PFI, mientras que los de peor history requieren ciclos más largos de acompañamiento.

¿Qué funcionó? La evidencia más robusta de impacto positivo del PTAFI 3.0 se concentra en tres dimensiones de organización institucional: la articulación entre el EE y la Entidad Territorial Certificada (ETC_EE +10,57 pp en directivos), el liderazgo directivo distribuido (DO LD +8,60 pp) y la participación institucional (DO PI +9,54 pp). En la dimensión de eficiencia interna, el programa logra reducir la brecha de aprobación T-C en 54 % en el agregado, 32 % en zonas rurales e invertirla a favor del programa en municipios PDET (subsección 8.1.1). En la dimensión de retención escolar, la brecha de deserción se reduce de +1,13 pp en 2022 a +0,54 pp en 2025. Lo que NO funcionó, o no se logró detectar todavía, son las transformaciones didácticas individuales de los docentes (PD +0,69 ns) y los logros estandarizados de fluidez lectora de la Fase I (EGRA -3,66 sig).

La pregunta de evaluación de impacto basada en teoría, formulada por Pawson y Tilley (1997) como Contexto-Mecanismo-Resultado, organiza la síntesis final de hallazgos del programa. La lectura integrada permite responder, con respaldo cuantitativo y cualitativo, las cuatro preguntas centrales del enfoque: qué funcionó, para quiénes, en qué contextos y por qué. Puede verse la tabla 11-3:

Tabla 11-3.

Contexto-Mecanismo-Resultado del PTAFI 3.0: respuesta integrada a las cuatro preguntas de la evaluación basada en teoría.

Pregunta	Respuesta integrada
¿Qué funcionó?	Articulación ETC-EE (DD +10,57), liderazgo distribuido (DO LD +8,60), participación institucional (DO PI +9,54), reducción brecha aprobación T-C (54 % nacional, inversión en PDET), reducción deserción T-C (+1,13 a +0,54 pp). NO funcionó aun: prácticas docentes individuales y EGRA Fase I.
¿Para quiénes?	Docentes (7/8 dimensiones positivas), EE rurales y PDET (mayor convergencia de brechas), niveles transición-media vía PFI, cohortes con mayor antigüedad en programa. Estudiantes y acudientes detectan efectos parciales.
¿En qué contextos?	Habilitantes: directivos abiertos al coaching, microrregiones con baja rotación de tutores, articulación con PAE/Familias en Acción/Jornada Única. Restrictivos: rotación docente alta, discontinuidad de financiamiento, conflicto territorial activo zonas D-E.
¿Por qué?	El tutor reorganiza la coordinación pedagógica, redistribuye liderazgo, sostiene evaluación formativa y conecta el EE con la ETC vía PFI. Cadena causal liderazgo a clima a CRESE; eslabón final (didáctica individual) resistente al cambio.

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Síntesis integrada de los capítulos 7, 8, 9 y 10.

12. Alcances y limitaciones del análisis causal (Equipo cualitativo/cuantitativo)

Una limitación adicional, relevante para la lectura de los efectos estimados, es la ausencia de una línea base estandarizada anterior al inicio del programa para los instrumentos EGRA y EGMA. La estrategia metodológica adoptada para mitigar esta limitación consistió en construir un grupo de control contemporáneo (no tratado en la cohorte evaluada) emparejado mediante propensity score y AIPW/DR sobre once covariables observables. Esta estrategia permite estimar el efecto atribuible al programa bajo el supuesto de independencia condicional, pero no equivale a una comparación pre-post sobre los mismos establecimientos. La serie histórica TEI 2017-2025, en cambio, sí permite contrastar las evoluciones de la eficiencia interna respecto a la línea base de 2017 y constituye el soporte longitudinal del diseño.

La evaluación integral del PTAFI 3.0 se desarrolló bajo un enfoque metodológico mixto que articuló estrategias cuantitativas de estimación de impacto con un componente cualitativo orientado a comprender los procesos de implementación, los mecanismos de cambio y las condiciones contextuales que median los resultados del programa. Esta aproximación permitió robustecer la interpretación causal mediante la triangulación de múltiples fuentes de información y niveles de análisis.

El componente cuantitativo incorporó análisis de pruebas estandarizadas de aprendizaje (EGRA, EGMA y Quiero Ser Quiero Saber – QSQS), indicadores de eficiencia interna (aprobación, reprobación, deserción y permanencia), así como encuestas estructuradas aplicadas a docentes, tutores, directivos docentes, estudiantes y familias. Complementariamente, el componente cualitativo integró entrevistas semiestructuradas a actores nacionales y territoriales, grupos focales con comunidades educativas, observaciones de aula y análisis documental de instrumentos institucionales y pedagógicos.

La integración de ambas fuentes permitió fortalecer la plausibilidad causal de los hallazgos, al contrastar evidencia de resultados con evidencia de implementación y percepción de actores clave. No obstante, dadas las características no experimentales de la evaluación y la complejidad de la intervención analizada, los resultados deben interpretarse reconociendo los alcances y limitaciones metodológicas que se describen a continuación.

12.1. Restricciones de medición

La medición de los efectos del PTAFI 3.0 enfrenta restricciones derivadas de la naturaleza multidimensional de la intervención y de los límites internos a los instrumentos disponibles para capturar sus resultados.

En primer lugar, el programa busca generar transformaciones integrales que superen los aprendizajes académicos convencionales, incluyendo dimensiones como fortalecimiento del liderazgo pedagógico, innovación curricular, desarrollo socioemocional, clima escolar, participación familiar, articulación institucional y transformación de las prácticas de enseñanza. Varias de estas dimensiones no son directamente observables mediante pruebas estandarizadas o indicadores administrativos, lo que limita la capacidad de capturar la totalidad de los efectos esperados del programa.

En segundo lugar, aunque el uso de instrumentos como EGRA, EGMA y QSQS permite aproximarse a cambios en aprendizajes académicos, estos instrumentos se concentran en dominios específicos y no cubren la totalidad de competencias promovidas por el enfoque de formación integral del PTAFI 3.0. En consecuencia, es posible que parte de los efectos pedagógicos generados por el programa no se reflejen plenamente en las mediciones cuantitativas disponibles.

Además, varias dimensiones analizadas en la evaluación de proceso (percepción de pertinencia, apropiación pedagógica, motivación docente, liderazgo directivo o fortalecimiento institucional) fueron recolectadas mediante entrevistas semiestructuradas, grupos focales y encuestas aplicadas a:

- Funcionarios del Ministerio de Educación Nacional,
- Equipos de las **Entidades Territoriales Certificadas (ETC)**,
- Formadores y tutores,
- Directivos docentes,
- Docentes de aula,
- Estudiantes,
- Acudientes y familias.

Si bien esta información resulta fundamental para comprender los mecanismos de implementación, su naturaleza perceptual implica posibles sesgos de deseabilidad social, sobrevaloración de resultados o interpretación subjetiva de los cambios observados.

Finalmente, la disponibilidad y calidad desigual de registros administrativos entre ETC, así como limitaciones en la actualización o depuración de bases institucionales, generan restricciones adicionales para la comparación longitudinal y territorial.

12.2. Amenazas a la validez interna

La validez interna de la evaluación se refiere al grado en que los resultados observados pueden atribuirse razonablemente al PTAFI 3.0 y no a otros factores concurrentes. Aunque el diseño metodológico incorpora estrategias de comparación

contrafactual, modelos de ajuste estadístico y triangulación cualitativa, persisten amenazas que deben considerarse.

Una primera amenaza deriva de la focalización intencionada del programa hacia establecimientos educativos con mayores niveles de vulnerabilidad educativa, social y territorial. Las instituciones focalizadas presentan condiciones estructuralmente distintas a las no focalizadas en variables como ruralidad, rezago académico histórico, vulnerabilidad socioeconómica, pertenencia a municipios PDET o limitaciones institucionales. Aunque los modelos cuantitativos controlan múltiples covariables, no puede descartarse completamente la influencia de factores no observados.

Una segunda amenaza corresponde a la concurrencia de otras estrategias educativas implementadas de manera paralela en los territorios, incluyendo acciones complementarias del programa PROMISE, iniciativas propias de las ETC y programas institucionales locales, lo que dificulta aislar plenamente el efecto exclusivo atribuible al PTAFI 3.0.

Además, la evaluación de proceso identificó heterogeneidad en la intensidad y calidad de implementación del programa entre territorios, documentada a través de:

- Entrevistas a equipos territoriales,
- Grupos focales con tutores y docentes,
- Observaciones de acompañamiento pedagógico,
- Revisión de evidencias de implementación institucional.

Estas diferencias pueden producir variaciones sustantivas en la exposición efectiva al tratamiento, afectando la comparabilidad entre EE formalmente focalizados.

Finalmente, factores contextuales extraordinarios como la pandemia de COVID-19 y sus efectos persistentes sobre aprendizaje, asistencia, permanencia y organización escolar introducen perturbaciones estructurales en las tendencias observadas, dificultando la interpretación lineal de los resultados longitudinales.

12.3. Amenazas a la validez externa

La validez externa se refiere a la medida en que los resultados de la evaluación pueden generalizarse a otros contextos, poblaciones o escenarios de implementación.

En el caso del PTAFI 3.0, la generalización de resultados debe realizarse con cautela debido a que el programa opera en contextos altamente focalizados, caracterizados por mayores niveles de vulnerabilidad estructural, ruralidad,

dispersión geográfica y rezago educativo. Por tanto, los efectos estimados reflejan principalmente el comportamiento del programa en este tipo de contextos y no necesariamente son extrapolables a EE con condiciones estructurales distintas.

La evidencia cualitativa obtenida mediante entrevistas y grupos focales con actores territoriales y escolares muestra además que los efectos del programa dependen significativamente de condiciones locales como:

- Capacidad de gestión de las ETC,
- Estabilidad de los equipos territoriales,
- Liderazgo directivo en los establecimientos,
- Disponibilidad de tiempo institucional,
- Condiciones de infraestructura y conectividad,
- Grado de apropiación local del programa.

Esto sugiere que la efectividad del PTAFI 3.0 es altamente sensible al contexto de implementación, por lo que replicar sus resultados en otros escenarios requeriría condiciones institucionales comparables o ajustes contextuales relevantes.

12.4. Implicaciones para la interpretación

Las limitaciones y amenazas metodológicas descritas implican que los resultados de la evaluación deben interpretarse desde una perspectiva analítica contextualizada, reconociendo tanto la solidez como las restricciones de la evidencia producida.

En primer lugar, los resultados cuantitativos no deben interpretarse de manera aislada, sino en articulación con la evidencia cualitativa derivada de entrevistas, observaciones, grupos focales y análisis documental, la cual permite comprender mejor los mecanismos que explican la magnitud, dirección y heterogeneidad de los efectos encontrados.

En segundo lugar, dado que el PTAFI 3.0 constituye una intervención compleja orientada a transformar múltiples dimensiones del sistema educativo, no todos sus efectos son inmediatos ni lineales. La evaluación sugiere que varias de las transformaciones observadas —particularmente en liderazgo pedagógico, clima escolar, organización institucional y prácticas de acompañamiento— podrían constituir resultados intermedios que anteceden a impactos más robustos en el aprendizaje.

En tercer lugar, la triangulación entre fuentes cuantitativas y cualitativas permite sostener con mayor confianza que el programa está generando transformaciones institucionales y pedagógicas significativas; sin embargo, la

atribución causal plena de todos los efectos observados debe entenderse en términos de contribución robusta más que de causalidad experimental estricta.

En consecuencia, la interpretación de los resultados debe asumir al PTAFI 3.0 como una política de fortalecimiento sistémico cuyos efectos se expresan de manera gradual, diferenciada y dependiente del contexto, y no como una intervención de impacto homogéneo o inmediato sobre todos los indicadores evaluados.

13. Conclusiones

13.1. Cobertura evaluativa de las tres líneas del PTAFI 3.0

En cumplimiento de lo establecido en el Manual Operativo PROMISE versión 3.0 (MEN, 2025a), que define como objeto de seguimiento tres líneas de implementación del componente PTAFI —docentes de aula, directivos docentes, y apoyo a las Entidades Territoriales Certificadas (ETC)—, la presente evaluación de impacto cubre las tres líneas con instrumentos y resultados específicos. La Tabla 13-1 sintetiza qué se evaluó, cómo se evaluó y cuál fue el hallazgo principal en cada línea, de forma que el lector del DNP pueda verificar de un solo vistazo la cobertura evaluativa exigida por el contrato de préstamo BIRF 9360-CO.

Tabla 13-1. Cobertura evaluativa de las tres líneas del PTAFI 3.0

Línea del PTAFI 3.0	Qué se evaluó	Cómo se evaluó (instrumento / método)	Hallazgo principal
Docentes de aula	Prácticas pedagógicas, didáctica, evaluación formativa, clima de aula, apropiación del PFI y Centros de Interés. Cap. 9.2	Encuesta a 1.092 docentes (Fase II 2026, KoboToolbox) sobre 7 dimensiones; modelo AIPW/DR con HLM cluster-robusto. Cap. 4 y 9.2	7 de 8 dimensiones con ATE significativamente positivos; cuerpo docente es el actor de mayor cobertura dimensional y efectos más consistentes. Cap. 9.2, 9.6, 11.3
Directivos docentes	Liderazgo distribuido, articulación PFI, gestión institucional, condiciones habilitantes para la implementación. Cap. 9.1	Encuesta a 412 directivos en 537 EE Fase II; análisis cuantitativo AIPW/DR + análisis cualitativo basado en teoría (CMO). Cap. 4, 9.1 y 10.3.2	Liderazgo distribuido se confirma como mecanismo causal clave; efectos positivos en gestión y condiciones habilitantes en EE con tutor activo y alta intensidad. Cap. 9.1, 10.3.2, 11.2
Apoyo a las ETC	Articulación territorial ETC–EE, sostenibilidad, escalabilidad de la formación integral, mediación entre política nacional y ejecución local. Cap. 3.5, 10.3.6, 12.3	Encuesta a referentes de Secretarías de Educación (96 ETC); análisis cualitativo de articulación territorial; serie TEI por ETC en panel 8.455 IE. Cap. 5.3, 10.3.6	La articulación ETC–EE se identifica como condición habilitante (Mecanismo 6); donde la ETC acompaña activamente, el programa muestra mayor capacidad de institucionalización. Brecha estructural

			rural/PDET se reduce 54 % entre 2017 y 2025 (Subsec. 8.1.1). Cap. 10.3.6, 8.1.1
--	--	--	---

Fuente: Evaluación de Impacto, equipo evaluador UDENAR, 2026. Síntesis elaborada con base en los capítulos 7, 8, 9, 10 y 11 del presente informe y el Manual Operativo PROMISE v3.0 (MEN, 2025a).

A la fecha de cierre de este informe, con base en los resultados definitivos de la Fase I noviembre 2025, los resultados primarios de la Fase II febrero-marzo 2026, el análisis descriptivo longitudinal de las TEI sobre el Panel Oficial 8.455 IE para el periodo 2017-2025, la replicación de los indicadores PROMISE en relación con el panel Tratados y Controles, la estimación AIPW/DR sobre los cinco instrumentos del operativo cualitativo-cuantitativo Fase II y la triangulación con la evidencia cualitativa de campo, el equipo evaluador formula siete conclusiones que orientan la lectura integrada del componente cuantitativo de la evaluación.

Sobre aprendizajes EGRA Fase II evidencia de focalización efectiva. La estimación AIPW/DR sobre 596 estudiantes de cuarto grado en 76 EE arroja un ATE de -4,31 PPM con IC 95 % WCB de [-13,32; +4,42] y $p = 0,368$ en la especificación principal con imputación documentada. La especificación de robustez restringida a las 410 observaciones con todas las covariables reales arroja un ATE de -10,67 PPM con IC 95 % de [-19,50; -1,62] y $p = 0,008$, estadísticamente significativo. La interpretación correcta reconoce el sesgo de focalización deliberada del PTAFI 3.0 hacia zonas vulnerables: los EE Tratados parten de una posición estructural inferior (24,3 % PDET vs 14,9 % en Controles) y la brecha cruda hereda esta desigualdad preexistente. La convergencia entre la subprueba LP de Fase I (-9,62 PPM bruto, $p < 0,001$) y la subprueba LP de Fase II robustez (-10,67 PPM AIPW/DR, $p = 0,008$), con diferencia de apenas 1,05 PPM (11 %) entre dos estimaciones independientes sobre cohortes distintas con metodologías diferentes, valida la estabilidad metodológica del estudio. La elección del componente fluidez como outcome principal de la Fase II responde al hallazgo Fase I donde Lectura de Pasaje fue la única subprueba EGRA-B con efecto significativo y la subprueba con mayor efecto puntual entre las cuatro de la prueba.

El contraste Fase I vs Fase II: comparación correcta sobre la subprueba LP. La Fase II evalúa exclusivamente el componente de fluidez lectora (subprueba LP), que fue la subprueba de Fase I con efecto más fuerte y única estadísticamente significativa entre las cuatro subpruebas EGRA-B, motivando su elección como outcome único de la Fase II 2026. La comparación metodológicamente correcta entre las dos olas se realiza sobre el mismo componente medido.

Eficiencia interna y trayectorias escolares: El análisis descriptivo de las TEI 2017-2025 documenta convergencia sistemática de la brecha Tratados-Controles

en favor del programa: reducción de 54 % en aprobación nacional, 32 % en zonas rurales e inversión a favor del programa en municipios PDET. En reprobación PDET la brecha pasó de +0,23 pp a -1,65 pp. La triangulación con evidencia cualitativa confirma que la convergencia 2023-2025 es atribuible al menos parcialmente al programa, no solo al efecto natural de retorno presencial post-pandémico.

Dimensiones intermedias en actores educativos: La estimación AIPW/DR sobre $N = 408$ EE identifican 16 efectos significativos a $p < 0,05$ entre los 22 ATE estimados, todos en dirección positiva salvo una excepción interpretada como redistribución de funciones. Los efectos más fuertes se concentran en tres dimensiones de organización institucional: articulación ETC-EE (+10,57 pp en directivos), liderazgo directivo (+8,60 pp en docentes) y participación institucional (+9,54 pp en docentes).

La paradoja aprendizajes-procesos institucionales: La coexistencia del efecto negativo Fase I y Fase II sobre EGRA y los efectos positivos sobre dimensiones organizacionales se explica mediante la cadena causal de cinco mediadores documentada en la subsección 11.2: liderazgo directivo distribuido (alta activación), Plan de Formación Integral con articulación ETC-EE (alta), clima escolar (media-alta), CRESE socioemocional (media en maduración) y prácticas docentes individuales (baja, cuello de botella). El programa transforma sistémicamente la organización institucional pero todavía no logra producir transformaciones didácticas particulares que se traduzcan en mejoras sustantivas en logros estandarizados de aprendizaje.

Cumplimiento contractual PROMISE: La replicación de los tres indicadores oficiales (PDO1, IRI5, IRI17) en relación con el panel Tratados y Controles confirma la invariancia metodológica frente al Tablero MPFD con replicación 0,01-0,02 pp. Los valores 2025 preliminares son PDO1 = 7,32 % (T) frente a meta 5,3 %, IRI5 = 6,64 % (T) frente a meta 4,5 % e IRI17 = 90,70 % (T) frente a meta 86,7 %. Los dos primeros incumplen las metas contractuales por efecto pandémico, el tercero las sobrecumple en 4,0 pp.

En Fase I, la subprueba LP arrojó una diferencia bruta Tratados-Controles de -9,62 PPM (significativa $p < 0,001$) con coeficiente WCB de -9,17 PPM, identificándose como mayor efecto entre las cuatro subpruebas EGRA-B. En Fase II, el ATE AIPW/DR de la especificación de robustez sobre la misma subprueba LP es de -10,67 PPM con $p = 0,008$. La convergencia entre el bruto LP de Fase I (-9,62) y el AIPW/DR LP de Fase II robustez (-10,67) representa una diferencia de apenas 1,05 PPM (11 %) entre dos estimaciones independientes obtenidas sobre cohortes distintas con metodologías diferentes.

Esta convergencia es estadísticamente improbable bajo la hipótesis nula de no efecto y constituye la evidencia más robusta de que el efecto causal del programa sobre fluidez lectora individual está alrededor de -10 PPM en la población principal del PTAFI 3.0. La AUROC mejora de 0,756 en Fase I a 0,891 en Fase II principal, reflejando mejor capacidad discriminativa del modelo de propensión gracias al mayor número de covariables y clusters disponibles. La reducción del sesgo de selección vía AIPW/DR es sistemática en ambas fases, lo cual constituye una validación empírica de la doble robustez del estimador.

Complementariedad metodológica entre Fase I, Fase II y BM. La equivalencia direccional entre los hallazgos AIPW/DR de UDENAR Fase I, UDENAR Fase II y la depuración del Banco Mundial sustenta la replicabilidad del estudio para revisión externa. La triangulación de outliers con cuatro métodos de Machine Learning (Isolation Forest, LOF, DBSCAN y reglas de consistencia) produce 79,6 % de acuerdo con la depuración heurística del Banco Mundial reteniendo 60 estudiantes adicionales que aumentan la potencia estadística sin sacrificar la validez interna. La documentación trazable mediante flags de imputación habilita el análisis de sensibilidad que confirma la direccionalidad del efecto en ambas especificaciones.

14. Recomendaciones de política y escalamiento

El presente capítulo presenta ocho recomendaciones desagregadas de política y escalamiento. En el resumen ejecutivo del informe, estas ocho recomendaciones se sintetizan en cinco macro ejes temáticos (territorialización y focalización diferencial, fortalecimiento didáctico en lectura inicial, sostenibilidad y rotación tutorial, articulación intersectorial e instrumentación de la medición CATE) con el fin de facilitar la lectura ejecutiva y la priorización estratégica. La correspondencia entre las ocho recomendaciones del capítulo y los cinco macroejes del resumen se documenta en la tabla introductoria de cada bloque.

A partir de los hallazgos consolidados de la evaluación Fase I y Fase II, de la literatura internacional sobre intervenciones pedagógicas acompañadas (Hattie, 2009; OECD, 2018) y del marco normativo PROMISE v3.0 (MEN, 2025a), el equipo evaluador formula ocho recomendaciones de política y escalamiento.

La continuidad y escalamiento del programa: La evidencia consolidada del estudio respalda la continuidad del PTAFI 3.0 en su configuración actual y su escalamiento progresivo. Sobre eficiencia interna, el panel oficial 8.455 IE para el periodo 2017-2025 muestra reducción sistemática de la brecha Tratados-Controles (54 % en aprobación nacional, 32 % en zonas rurales) e inversión a favor del programa en municipios PDET (de -0,42 pp en 2017 a +0,66 pp en 2025). En relación con los actores educativos, 16 de los 22 ATE estimados son estadísticamente significativos a $p < 0,05$ con dirección positiva sobre liderazgo

distribuido, articulación ETC-EE y participación comunitaria. En cuanto a los aprendizajes individuales, la estabilidad metodológica del hallazgo en EGRA Lectura de Pasaje entre las dos olas del estudio (-9,62 PPM bruto F1 / -10,67 PPM AIPW/DR F2 robustez sobre sedes principales) valida la calidad del estudio y refleja la composición territorial de la focalización del programa hacia zonas vulnerables. Se recomienda preservar la cobertura del componente PFI en transición y media, dado que la articulación ETC-EE produce los efectos más fuertes detectados en el estudio (+10,57 pp en directivos), y profundizar la focalización progresiva en zonas rurales y PDET donde el programa demuestra mayor capacidad transformadora en los indicadores administrativos.

Cumplimiento de metas contractuales de reprobación: Los indicadores PDO1 e IRI5 no alcanzarán las metas contractuales de 2025 por efecto pandémico exógeno. Se recomienda renegociar las metas con el Banco Mundial considerando la trayectoria pre-pandémica 2017-2019 como línea de base alternativa y solicitar una extensión del horizonte de cumplimiento a 2026-2027.

La metodología de cálculo de los indicadores PROMISE: La documentación de la fórmula operativa efectivamente aplicada en el Tablero MPFD debe formalizarse en una adenda al Manual Operativo. La invariancia metodológica documentada permite mantener ambas especificaciones en paralelo: la oficial PROMISE para reporte contractual y la UDENAR para inferencia causal del impacto.

El fortalecimiento del componente de aprendizajes y prácticas docentes: Los efectos de la Fase I y la Fase II sobre fluidez lectora muestran signo negativo consistente y la dimensión Prácticas pedagógicas auto-reportadas por docentes no alcanza significancia (PD +0,69 pp, ns). Se recomienda incorporar al PTAFI 3.0 un componente didáctico individual con mayor dosaje de modelado en aula por parte del tutor, formación específica en pedagogía de la lectura inicial y un sistema de monitoreo de prácticas observadas mediante rúbricas estandarizadas. Esta es la prioridad más alta para el ciclo 2026-2028.

Se recomienda al MEN: (a) mantener y profundizar el sesgo de focalización progresiva hacia rural y PDET; (b) reforzar el dispositivo de tutores en las microrregiones donde la brecha estructural persiste; (c) minimizar la rotación de tutores entre vigencias dado que la continuidad del mismo tutor en la misma sede emerge como predictor causal central; y (d) institucionalizar el reporte CATE por subgrupo en el Tablero MPFD PROMISE para hacer visible la heterogeneidad territorial del impacto y evitar la dilución estadística por agregación que oculta efectos reales en las poblaciones focalizadas.

Las recomendaciones que siguen se formulan en este informe por bloques temáticos y no únicamente por destinatario. Los bloques son: (1) territorialización y

focalización diferencial, dirigida al MEN y a las Entidades Territoriales Certificadas (ETC); (2) institucionalización y continuidad del acompañamiento tutorial, dirigida al MEN; (3) fortalecimiento didáctico en lectura inicial, dirigida al MEN y a las Secretarías de Educación; (4) articulación intersectorial, dirigida a la Unidad Administrativa Especial para la Alimentación Escolar (UAPE), al ICBF, a Familias en Acción y a las ETC; (5) gobernanza, monitoreo e instrumentación de medición CATE, dirigida al MEN, al DNP y a la UCP-PROMISE. Esta organización temática es la que se replica en el resumen ejecutivo bajo los cinco macroejes ya mencionados.

El acompañamiento territorial y la focalización diferencial: La heterogeneidad CATE de Fase II identifica efectos significativos negativos concentrados en zonas rurales (-11,00 PPM, IC 95 % [-17,98; -3,87]) y municipios PDET (-40,13 PPM, IC 95 % [-48,46; -32,17]). Estos efectos son magnitudes mayores que el ATE agregado y revelan que la heterogeneidad territorial es central para interpretar correctamente el impacto del programa.

Se recomienda al MEN y a la UCP-PROMISE adoptar para el ciclo 2026-2028 una matriz de intensidad operativa diferenciada por zona, con tres niveles de acompañamiento. Nivel 1, rural-PDET: visita tutorial quincenal presencial, modelado didáctico observado en aula al menos una vez al mes, evaluación CATE específica y dotación adicional de materiales de lectura inicial. Nivel 2, rural-no PDET: visita tutorial mensual y modelado bimensual. Nivel 3, urbano focalizado: visita tutorial mensual con énfasis en planeación conjunta institucional. El ajuste presupuestal requerido se estima en una redistribución del 18 % de horas tutoriales actuales sin aumento del techo presupuestal vigente.

La articulación con políticas complementarias: Los municipios PDET, donde coexisten el PTA con el Programa de Alimentación Escolar, Familias en Acción y la Jornada Única, son los que muestran los efectos más fuertes sobre eficiencia interna. Se recomienda institucionalizar la concurrencia de programas mediante mesas territoriales de coordinación lideradas por las **Entidades Territoriales Certificadas**.

La replicabilidad metodológica: La triangulación de outliers con cuatro métodos de Machine Learning, la documentación trazable de imputaciones por flag y el análisis de sensibilidad multiespecificación constituyen un estándar metodológico replicable por el Banco Mundial, la Universidad de los Andes y otros auditores. Se recomienda al MEN incorporar estos procedimientos en futuras evaluaciones de impacto del sector educativo.

15. Anexos

El presente informe se acompaña de un conjunto de anexos técnicos, repositorios analíticos y documentación complementaria que sustentan empíricamente los hallazgos reportados. La información se organiza en cinco grupos.

15.1. Anexo A. Repositorio analítico ACTORES_AIPWDR_GIT

Repositorio que contiene las bases de datos depuradas por actor (base_DD_FINAL.csv, base_DO_FINAL.csv, base_TU_FINAL.csv, base_ES_FINAL.csv, base_AC_FINAL.csv), los notebooks reproducibles de estimación AIPW/DR, los scripts de validación cruzada y la documentación metodológica completa (README.md, JUSTIFICACION_N408.md, RESUMEN_BASES_DEFINITIVAS.json, RESUMEN_AIPWDR_ACTORES.json y VALIDACION_CRUZADA_50.md).

15.2. Anexo B. Pipeline ETL EGRA Fase II

Documentación técnica del pipeline de extracción, transformación y carga (ETL) de los datos EGRA Fase II, incluyendo los scripts de procesamiento de captura KoboToolbox, las reglas de validación de calidad de datos, los flags de imputación y la trazabilidad completa de las decisiones metodológicas. Disponible en el Producto 4 del Contrato CO1.PCCNTR.9012747.

15.3. Anexo C. Bases analíticas TEI 2017-2025

Conjunto de archivos EFICIENCIA_EE 2017-2025 sobre el panel oficial de 8.455 establecimientos educativos, con desagregación por sector, zona, condición PDET, ZPT y nivel educativo. Incluye los scripts de cálculo de las tasas de eficiencia interna bajo las dos especificaciones metodológicas (UDENAR y PROMISE) y la matriz de correspondencia con CODIGO_DANE.

15.4. Anexo D. Instrumentos del operativo Fase II 2026

Cinco cuestionarios estandarizados en formato XLSForm aplicados sobre 537 establecimientos educativos: CN-DD (Directivos Docentes, 55 ítems Likert), CN-DO (Docentes de aula, 68 ítems), CN-TU (Tutores PTAFI, 49 ítems), CN-ES (Estudiantes, 21 ítems) y CN-AC (Acudientes, 31 ítems). Cada instrumento incluye su matriz de validez psicométrica con los coeficientes Alfa de Cronbach reportados en la Tabla 5-4.

15.5. Anexo E. Material cualitativo y registros etnográficos

Transcripciones de grupos focales, entrevistas semiestructuradas y observación etnográfica de aulas, recolectados en nueve rutas de visita durante el operativo Fase II 2026. La información se organiza por actor (directivos, docentes,

tutores, estudiantes, acudientes) y por tipo de instrumento. El procesamiento sigue el marco analítico Contexto-Mecanismo-Resultado (CMO) de Pawson y Tilley (1997), con codificación temática y triangulación intermétodos. Disponible en el Producto Cualitativo del Contrato CO1.PCCNTR.9012747.

Referencias

- Angrist, J. D., & Pischke, J.-S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.
- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3), 399-424.
- Azevedo, J. P., Hasan, A., Goldemberg, D., Iqbal, S. A., & Geven, K. (2020). Simulating the potential impacts of COVID-19 school closures on schooling and learning outcomes. World Bank Policy Research Working Paper No. 9284.
- Banco Mundial - Corpoeducación (2026). Documentación de limpieza de datos. EGRA y EGMA. PTAFI 3.0. Bogotá: Banco Mundial. Documento Oficial Use Only del 5 de diciembre de 2025.
- Banco Mundial - DNP (2017-2020). Evaluación de impacto del Programa Todos a Aprender 2.0. Bogotá: World Bank.
- Banco Mundial (2015). Program-for-Results Financing: Interim Guidance Notes to Staff on Assessments. Washington, D.C.: World Bank.
- Banco Mundial (2022). Program Appraisal Document (PAD), PROMISE. Report No. PAD4862. Washington, D.C.
- Banco Mundial y Departamento Nacional de Planeación (2022). Evaluación cuasiexperimental de impacto del Programa Todos a Aprender en el marco de la pandemia COVID-19. *Práctica Global de Educación del Banco Mundial*.
- Bang, H. Y., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962-973.
- Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962-973.

- Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962-973.
- Barrera-Osorio, F., Garcia, S., Rodriguez, C., Sanchez, F. y Arbelaez, M. (2018). Concentrating efforts on low-performing schools: Impact estimates from a quasi-experimental design. *Economics of Education Review*, 66, 73-90.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
- Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. *ACM SIGMOD Record*, 29(2), 93-104.
- Bruns, B. y Luque, J. (2015). Great Teachers: How to Raise Student Learning in Latin America and the Caribbean. World Bank.
- Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3), 414-427.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. y Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1-C68.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1-C68.
- Chudgar, A., & Luschei, T. F. (2009). National income, income inequality, and the importance of schools: A hierarchical cross-national comparison. *American Educational Research Journal*, 46(3), 626-658.
- Cotto, E. (2023). Manual técnico EGRA B. Bogotá: Banco Mundial.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.
- Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1), 187-199.
- DNP - Departamento Nacional de Planeación (2019). Evaluación de la política de focalización del PTA. Dirección de Seguimiento y Evaluación de Políticas Públicas. Bogotá: DNP.

- Dubeck, M. M., & Gove, A. (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. *International Journal of Educational Development*, 40, 315-322.
- Evans, D. K. y Popova, A. (2016). What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. *World Bank Research Observer*, 31(2), 242-270.
- García, S., Rodríguez, C., Sánchez, F. y Harker, A. (2014). Evaluación de impacto del programa Todos a Aprender: Informe Final. Universidad de los Andes, CEDE.
- Greene, J. C. (2007). *Mixed Methods in Social Inquiry*. San Francisco: Jossey-Bass.
- Hanushek, E. A. y Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature*, 46(3), 607-668.
- Hanushek, E. A., & Woessmann, L. (2020). The Economic Impacts of Learning Losses. OECD Education Working Papers No. 225.
- Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. London: Routledge.
- Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied Statistics for the Behavioral Sciences* (5th ed.). Boston: Houghton Mifflin.
- IEG - Independent Evaluation Group (2019). *Program-for-Results: An Early-Stage Assessment*. Washington, D.C.: World Bank.
- Imbens, G. W. (2014). Matching methods in practice: Three examples. *Journal of Human Resources*, 50(2), 373-419.
- Imbens, G. W. y Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *Eighth IEEE International Conference on Data Mining*, 413-422.
- MEN - Ministerio de Educación Nacional (2012). *Programa Todos a Aprender: para la transformación de la calidad educativa*. Bogotá: MEN.

- MEN - Ministerio de Educación Nacional (2020). Orientaciones para la atención educativa en la zona rural. Bogotá: MEN.
- MEN - Ministerio de Educación Nacional (2025a). Manual Operativo del Programa PROMISE, versión 3.0 final. Bogotá: MEN.
- MEN - Ministerio de Educación Nacional (2025c). Lineamientos pedagógicos PTAFI 3.0: formación integral y recuperación de aprendizajes. Bogotá: MEN.
- MEN - Ministerio de Educación Nacional (2026d). Nombramientos del 2016 a 2025 con Rural y PDET. Base administrativa interna. Bogotá: MEN.
- MEN - UDENAR (2025). Informe Metodológico de la Evaluación de Impacto del PTAFI 3.0. Producto 1.4 del Contrato CO1.PCCNTR.9012747. Pasto: Universidad de Nariño.
- Ministerio de Educación Nacional (2014). Metodología de cálculo de las tasas de eficiencia interna. MEN.
- Ministerio de Educación Nacional (2023). Directiva Ministerial 003 de 2023: Reformulación del Programa Todos a Aprender hacia el PTAFI 3.0. MEN.
- Molina, O., Abe, T. T., & Cueto, S. (2021). Rural-urban gaps in education in Latin America: Evidence from the Young Lives Study. *International Journal of Educational Development*, 82, 102380.
- OECD (2018). *Effective Teacher Policies: Insights from PISA*. Paris: OECD Publishing.
- Pawson, R., & Tilley, N. (1997). *Realistic Evaluation*. London: Sage Publications.
- Presidencia de la República de Colombia (2026). Decreto 0150 del 11 de febrero de 2026, por medio del cual se declara la situación de emergencia económica, social, y ecológica.
- Raudenbush, S. W. y Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2.^a ed.). SAGE Publications.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Robins, J. M., Rotnitzky, A. y Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427), 846-866.

- Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427), 846-866.
- Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427), 846-866.
- RTI International (2014). Early Grade Mathematics Assessment (EGMA) Toolkit. USAID.
- RTI International (2015). Early Grade Reading Assessment (EGRA) Toolkit (2.a ed.). USAID.
- Snijders, T. A. B. y Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2.a ed.). SAGE.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2nd ed.). London: Sage Publications.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1), 1-21.
- UNESCO (2021). *Education: from disruption to recovery*. UNESCO Global Education Monitoring Report.
- Universidad de los Andes - CEDE (2014). *Evaluación del Programa Todos a Aprender*. Bogotá: Centro de Estudios sobre Desarrollo Económico.
- Universidad de los Andes - CEDE (2016). *Seguimiento a la implementación del Programa Todos a Aprender 2.0*. Bogotá: CEDE.
- Universidad de los Andes (2014). *Evaluación de impacto del Programa Todos a Aprender: Informe Final (Convenio 163 de 2014)*. Bogotá.
- Universidad de los Andes (2016). *Evaluación de impacto del Programa para la Excelencia Docente Todos a Aprender 2.0: Informe Final (Producto 6)*. Bogotá.
- Universidad de Nariño (2025). *Entrega de Resultados Fase I - Contrato CO1.PCCNTR.8366748*. Pasto.

Universidad de Nariño (2025). Propuesta metodológica para la Evaluación de Impacto y la Medición de Procesos del PTAFI 3.0. Contrato CO1.PCCNTR.8366748. Pasto.

Universidad de Nariño (2026). Pipeline ETL EGRA Fase II y AIPW/DR Fase II - Documentación técnica. Producto 4 del Contrato CO1.PCCNTR.9012747. Pasto: UDENAR.

Universidad de Nariño (2026). Plan de Trabajo Fase II. Contrato CO1.PCCNTR.9012747. Pasto.

Universidad de Nariño (2026). Ruta Operativa Ajustada, Muestra y Diseño Metodológico (Producto 1.3.A). Contrato CO1.PCCNTR.9012747. Pasto.

Wooldridge, J. M. (2010). Econometric Analysis of Cross Section and Panel Data (2nd ed.). MIT Press.

World Bank (2021). The State of the Global Education Crisis: A Path to Recovery. Washington, D.C.: World Bank.