Educación | Icfes

Notas de Investigación



Trayectorias de educación inicial y desempeño en la prueba Saber 3°: construcción y potencial analítico de una base integrada de SIMAT, ICBF y Saber

Investigadores(as) del proyecto

Miguel Andrés Serrato Cáceres
Juanita Cifuentes González
Silvia Consuelo Gómez Soler
Jorge Leonardo Rodríguez-Arenas

Agosto 2026

Notas de Investigación
Edición No. 26
Agosto, 2026
Edición digital

Instituto Colombiano para la Evaluación de la Educación
Oficina de Gestión de Proyectos de Investigación
Calle 26 N. 69-76, Edificio Elemento, Torre II, piso 18**,** Bogotá, D. C.
Teléfono: (601) 4841410
proyectosinvestigacion@icfes.gov.co
https://www.icfes.gov.co

Directora general
Elizabeth Blandón Bermúdez

Jefe de Oficina de Gestión de Proyectos de Investigación
Paola Guio Veloza

Autores

Jorge Leonardo Rodríguez-Arenas
Miguel Andrés Serrato Cáceres
Juanita Cifuentes González
Silvia Consuelo Gómez Soler

Advertencia:
El contenido de este documento es el resultado de investigaciones y obras protegidas por la legislación nacional e internacional. No se autoriza su reproducción, utilización ni explotación a NINGÚN tercero. Solo se autoriza su uso para fines exclusivamente académicos. Esta información no podrá ser alterada, modificada o enmendada.

Citar este documento en estilo APA así:
Serrato Cáceres, M. A., Cifuentes González, J., Gómez Soler, S. C., & Rodríguez-Arenas, J. L. (2026*). Trayectorias de educación inicial y desempeño en la prueba Saber 3.º: Construcción y potencial analítico de una base integrada de SIMAT, ICBF y Saber* (Nota de investigación No. 26). Instituto Colombiano para la Evaluación de la Educación (Icfes). https://www.icfes.gov.co/web/guest/Nota-de-Investigación

Trayectorias de educación inicial y desempeño en la prueba Saber 3.º: construcción y potencial analítico de una base integrada de SIMAT, ICBF y Saber[*]

Miguel Andrés Serrato Cáceres[†] Juanita Cifuentes González[†]
Silvia Consuelo Gómez Soler[†] Jorge Leonardo Rodríguez-Arenas[‡]

Resumen

Esta nota documenta la construcción de una base de datos que vincula los registros de matrícula del Sistema Integrado de Matrícula (SIMAT), los registros de beneficiarios de programas de primera infancia del Instituto Colombiano de Bienestar Familiar (ICBF) y los resultados de la prueba Saber 3.º del Icfes. La base reconstruye la trayectoria de educación inicial de 33.007 estudiantes evaluados en 2022 y 2023, que representan a 821.721 estudiantes. Para cada uno, identifica si estuvo matriculado en algún grado de preescolar o fue beneficiario de un programa del ICBF durante los tres años previos a su ingreso a primero de básica primaria. La nota describe las fuentes, el procedimiento de vinculación y los criterios de depuración. También presenta evidencia descriptiva sobre el desempeño según el tipo de trayectoria registrada y precisa los alcances de esta nueva infraestructura de información. Los resultados muestran que los estudiantes sin registro administrativo de educación inicial se ubican en una posición intermedia de la distribución de desempeño.

Palabras clave: educación inicial, primera infancia, registros administrativos, vinculación de registros, trayectorias educativas, Saber 3.º, SIMAT, ICBF, Colombia.

Clasificación JEL: I21, I24, I28, C81.

[*] Agradecemos al Icfes por brindarnos el acceso a los datos en este artículo a través del proyecto de investigación interna "OFINVES-2025-18 Primera Infancia PUJ". Las opiniones contenidas en el presente documento son responsabilidad exclusiva de sus autores y no comprometen al Icfes ni a las entidades u organizaciones a las que estos pertenecen.
[†] Pontificia Universidad Javeriana, Bogotá, Colombia.
[‡] Oficina de Gestión de Proyectos de Investigación (OGPI), Icfes, Bogotá, Colombia.

Early Childhood Education Trajectories and Third-Grade Test Performance: Construction and Analytical Potential of an Integrated SIMAT, ICBF, and Saber Database

Miguel Andrés Serrato Cáceres Juanita Cifuentes González
Silvia Consuelo Gómez Soler Jorge Leonardo Rodríguez-Arenas

Abstract

This note documents the construction of a database that links enrollment records from Colombia's Integrated Enrollment System (SIMAT), beneficiary records from the early childhood programs of the Colombian Institute of Family Welfare (ICBF), and results from the Icfes Saber third-grade assessment. The database reconstructs the early childhood education trajectories of 33,007 students assessed in 2022 and 2023, who represent 821,721 students. For each student, it identifies whether the child was enrolled in a preschool grade or participated in an ICBF program during the three years preceding first grade. The note describes the data sources, the record linkage procedure, and the sample restrictions. It also presents descriptive evidence on performance by type of registered trajectory and clarifies the scope of this new information infrastructure. The results show that students without an administrative record of early childhood education occupy an intermediate position in the performance distribution.

Keywords: early childhood education, administrative records, record linkage, educational trajectories, Saber 3, SIMAT, ICBF, Colombia.

JEL Classification: I21, I24, I28, C81.

Contenido

1 Introducción **1**

2 Datos y vinculación de registros **2**

2.1 Prueba Saber 3.º **2**

2.2 Trayectoria de educación inicial: SIMAT y registros del ICBF **2**

2.3 Información municipal complementaria **3**

2.4 Composición de la base **3**

3 Resultados descriptivos **4**

3.1 Desempeño según tipo de trayectoria **4**

3.2 Diferencias ajustadas por características observables **6**

4 Alcances y limitaciones de la base **7**

5 Potencial analítico y agenda de investigación **8**

6 Conclusiones **9**

1. Introducción

En Colombia, la evidencia causal sobre los efectos de la educación inicial se concentra principalmente en resultados medidos durante la primera infancia. Los estudios disponibles (Bernal et al., 2009; Bernal, 2015; Attanasio et al., 2014; Bernal et al., 2019; Andrew et al., 2024) analizan el desarrollo cognitivo, nutricional y socioemocional mediante instrumentos aplicados antes del ingreso a la escuela. El seguimiento más prolongado identificado para el país llega hasta los cinco años y medio de edad (Andrew et al., 2018). Por esta razón, aún es limitada la evidencia sobre la relación entre las experiencias de educación inicial y las trayectorias escolares posteriores.

Esta limitación también se observa en otros países de ingreso bajo y medio. Una revisión sistemática reciente de intervenciones de cuidado infantil identifica setenta y un estudios, pero solo dos reportan resultados correspondientes a etapas posteriores de la vida (Evans et al., 2024). Una de las dificultades para realizar este tipo de seguimiento es que los sistemas que registran la atención a la primera infancia no siempre se encuentran articulados con los sistemas de información del sector educativo.

Colombia cuenta con fuentes administrativas que permiten avanzar en esa dirección. El Sistema Integrado de Matrícula (SIMAT) contiene registros censales y nominales de la matrícula escolar. El ICBF dispone de registros nominales de los beneficiarios de sus programas de primera infancia y el Icfes administra evaluaciones estandarizadas en diferentes momentos de la trayectoria escolar. Sin embargo, hasta donde conocemos, no se había documentado una vinculación de estas tres fuentes con fines de investigación.

Esta nota presenta el proceso de vinculación. A partir de la muestra de estudiantes evaluados en la prueba Saber 3.º en 2022 y 2023, se reconstruye la trayectoria de educación inicial registrada durante los tres años anteriores al ingreso esperado a primero de básica primaria. Para ello, se combina la información de SIMAT con los registros de beneficiarios del ICBF. La base resultante permite analizar, para una muestra representativa a escala nacional, la relación entre el tipo de trayectoria de educación inicial registrada y el desempeño en Saber 3.º.

La nota tiene tres objetivos. En primer lugar, documenta las fuentes, el procedimiento de vinculación y la composición de la base para facilitar la réplica y la extensión del ejercicio. En segundo lugar, presenta evidencia descriptiva sobre el desempeño según el tipo de trayectoria registrada. Finalmente, precisa los alcances y las limitaciones de la información disponible. Como se explica en la Sección 4, la base permite caracterizar las trayectorias con un nivel de detalle que no estaba disponible, pero no permite estimar por sí sola el efecto causal de la educación inicial sobre el aprendizaje.

El resultado descriptivo que orienta el análisis es que los estudiantes sin registro de matrícula preescolar ni de participación en programas del ICBF no se ubican en el extremo inferior de la distribución de desempeño. Por el contrario, ocupan una posición intermedia y

su promedio supera el de algunos grupos que registran atención en establecimientos oficiales. Este patrón no permite equiparar la ausencia de registro administrativo con la ausencia de atención. También es consistente con la posibilidad de que este grupo incluya, en una proporción que no puede determinarse con la información disponible, niños que recibieron formas de cuidado o educación inicial no observadas en las fuentes utilizadas.

2. Datos y vinculación de registros

2.1. Prueba Saber 3.º

La fuente principal de los resultados académicos es la prueba Saber 3.º, aplicada por el Icfes en 2022 y 2023. Esta evaluación hace parte de las pruebas Saber aplicadas en los grados 3.º, 5.º, 7.º y 9.º. En grado tercero se evalúan dos áreas: Lectura y Matemáticas.

El diseño de la prueba condiciona el tratamiento estadístico de los resultados. Desde el piloto de 2021, las aplicaciones de Saber en estos grados siguen un diseño *muestral* probabilístico, estratificado y multietápico. Este diseño permite realizar inferencias válidas para grupos de estudiantes, establecimientos, entidades territoriales o para el sistema en su conjunto, pero no permite evaluar con precisión el desempeño individual. En consecuencia, la unidad observada en la base es el estudiante evaluado y la unidad de inferencia es poblacional.

El análisis incorpora tres elementos del diseño de medición. Primero, los resultados de Lectura y Matemáticas se analizan a partir de los valores plausibles reportados por el Icfes. Estos valores no corresponden a puntajes individuales exactos, sino a realizaciones posibles de la habilidad latente del estudiante. Las estimaciones se realizan por separado para cada valor plausible y luego se combinan siguiendo la lógica de imputación múltiple, de manera que la varianza total incorpore la incertidumbre muestral y la asociada a la medición. Segundo, todas las estadísticas emplean los pesos muestrales finales. Tercero, los errores estándar incorporan las réplicas de pesos provistas por el diseño para reflejar la estratificación y la conglomeración de la muestra.

2.2. Trayectoria de educación inicial: SIMAT y registros del ICBF

La variable de interés se construye rastreando, para cada estudiante evaluado, su situación durante lo que en adelante se denomina la *etapa preescolar*. El Decreto 1075 de 2015 establece que el nivel preescolar se ofrece a educandos de tres a cinco años y comprende los grados de prejardín, jardín y transición. En consecuencia, la etapa preescolar de cada niño se define aquí como el periodo de tres años anteriores al momento en que debería encontrarse matriculado en primero de básica primaria en ausencia de repitencia. Para los estudiantes evaluados en 2022, ese periodo corresponde a los años 2016 a 2018. Para los evaluados en 2023, comprende los años 2018 a 2020.

Durante esa ventana se determina, mediante SIMAT, si el niño estuvo matriculado en algún grado de preescolar en establecimientos educativos oficiales o no oficiales, y mediante

los registros del ICBF, si fue beneficiario de alguno de los programas de primera infancia ofrecidos por esa entidad. La vinculación entre fuentes se apoya en el procedimiento estandarizado de emparejamiento de registros del Icfes, que enlaza individuos entre bases mediante criterios de coincidencia exacta construidos sobre combinaciones de número de documento de identidad, nombre completo y fecha de nacimiento, priorizando los criterios que usan el documento de identificación.

A partir de ese rastreo se define una variable binaria que toma el valor de uno si el niño registra matrícula en al menos un grado de preescolar o participación en un programa de primera infancia del ICBF durante la etapa preescolar. La variable toma el valor de cero si no aparece ninguno de estos registros. Dado que esta medida agrupa experiencias potencialmente heterogéneas, el análisis distingue dos dimensiones adicionales: la participación en programas del ICBF y la naturaleza oficial o no oficial del establecimiento educativo en el que se registró la matrícula. El cruce de ambas dimensiones define los cuatro grupos de trayectoria que organizan los resultados de la Sección 3.

Es importante subrayar el alcance de esta medida. Se trata de un registro administrativo de matrícula o de beneficio, no de una medida de asistencia efectiva ni de exposición acumulada. La información disponible no permite identificar la duración de la asistencia, la edad exacta de ingreso, la continuidad de la trayectoria, la intensidad de la participación, la calidad del servicio recibido ni las características pedagógicas del establecimiento o programa. Un niño que registra matrícula en un solo año de preescolar se clasifica igual que uno que cursó los tres grados, y un niño atendido en un Hogar Comunitario de Bienestar se clasifica igual que uno atendido en un Centro de Desarrollo Infantil.

2.3. Información municipal complementaria

La base se complementa con información contextual del municipio de nacimiento de cada niño durante su etapa preescolar. Esta información incluye las tasas de cobertura de los grados de preescolar, construidas a partir del censo de establecimientos educativos C-600 del DANE para el periodo entre 2014 y 2021. También incluye el número de beneficiarios de programas de primera infancia publicado por el Observatorio del Bienestar Familiar del ICBF para el periodo entre 2018 y 2020, así como el índice de necesidades básicas insatisfechas del Censo Nacional de Población y Vivienda de 2018. Estas variables caracterizan las condiciones de oferta que enfrentó cada niño y se utilizan para construir comparaciones ajustadas entre grupos de trayectoria.

2.4. Composición de la base

El ejercicio produce una base con 33.007 observaciones de estudiantes evaluados en Lectura y Matemáticas en 2022 y 2023. De estas observaciones, 16.540 corresponden a 2022 y 16.467 a 2023. De acuerdo con los pesos muestrales, la base representa a 821.721 estudiantes, 426.648 en 2022 y 395.073 en 2023. La muestra se restringe a estudiantes de

ocho años o menos al momento de la aplicación, con el propósito de excluir situaciones de extraedad y repitencia escolar. En la Sección 4 se discuten las consecuencias de esta restricción.

La Tabla 1 resume la composición resultante. De los 33.007 estudiantes, 31.266 registran alguna matrícula preescolar o participación en programas del ICBF durante su etapa preescolar, y 1.741 no registran ninguna de las dos.

TABLA 1: Composición de la base por tipo de trayectoria de educación inicial registrada

Grupo de trayectoria	Observaciones	% del total
Sin registro administrativo de educación inicial	**1.741**	**5,3**
Con registro administrativo	**31.266**	**94,7**
Con participación en programas del ICBF	21.051	63,8
En establecimiento educativo oficial	16.737	50,7
En establecimiento educativo no oficial	4.314	13,1
Sin participación en programas del ICBF	10.215	30,9
En establecimiento educativo oficial	4.528	13,7
En establecimiento educativo no oficial	5.687	17,2
Total	**33.007**	**100,0**

Nota: las observaciones corresponden a estudiantes evaluados en la prueba Saber 3.º en 2022 y 2023 con ocho años o menos al momento de la aplicación. Los porcentajes se calculan sobre el total de observaciones sin ponderar. *Fuente:* elaboración de los autores con información de la prueba Saber 3.º (Icfes), SIMAT (MEN) y registros de beneficiarios de primera infancia del ICBF.

El grupo sin registro representa el 5,3 % de las observaciones. Su tamaño, en comparación con los 31.266 estudiantes que tienen algún registro, reduce la información disponible para construir grupos comparables. La Sección 4 analiza las implicaciones de este desbalance.

3. Resultados descriptivos

3.1. Desempeño según tipo de trayectoria

La Tabla 2 presenta los promedios de Lectura y Matemáticas por grupo de trayectoria, calculados a partir de los valores plausibles, los pesos muestrales y las réplicas. En la comparación agregada, el grupo sin registro administrativo de educación inicial obtiene promedios superiores a los del conjunto de estudiantes con registro. Las diferencias son de 16,22 puntos en Lectura y 14,31 puntos en Matemáticas.

El promedio agregado oculta diferencias importantes entre tipos de trayectoria. El Gráfico 1 presenta estas diferencias en unidades de desviación estándar. Los estudiantes que participaron en programas del ICBF y estuvieron matriculados en establecimientos oficiales tienen los promedios más bajos. En Lectura, su promedio es 0,47 desviaciones estándar menor que el del grupo sin registro. En contraste, quienes no participaron en programas del ICBF y estuvieron matriculados en establecimientos no oficiales tienen un promedio 0,50

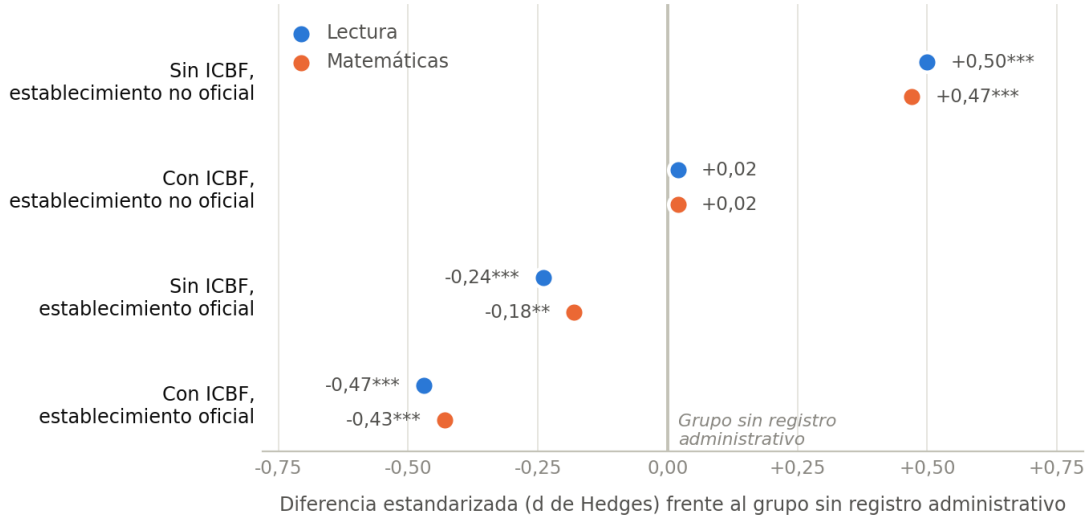
desviaciones estándar mayor. El grupo sin registro administrativo ocupa una posición intermedia y su promedio es similar al de los estudiantes que combinaron atención del ICBF con matrícula en establecimientos no oficiales.

TABLA 2: Promedios en Saber 3.º y diferencias frente al grupo sin registro administrativo

		Lectura		Matemáticas	
Grupo	N	Media	Diferencia	Media	Diferencia
Sin registro administrativo	1.741	423,49	n.a.	419,83	n.a.
Con registro administrativo	31.266	407,26	−16,22***	405,52	−14,31**
Con ICBF	21.051	394,79	−28,69***	393,64	−26,19***
Oficial	16.737	388,31	−35,18***	387,79	−32,04***
No oficial	4.314	425,37	+1,88	421,22	+1,38
Sin ICBF	10.215	436,80	+13,31***	433,66	+13,83**
Oficial	4.528	404,54	−18,95***	405,59	−14,25**
No oficial	5.687	462,71	+39,23***	456,21	+36,38***

Nota: el cálculo de medias incorpora valores plausibles, pesos muestrales finales y réplicas de pesos. Se integran las aplicaciones de 2022 y 2023 en un único conjunto de datos respetando el diseño muestral de cada una. La diferencia corresponde a una prueba de medias frente al grupo sin registro administrativo. * p < 0,10, ** p < 0,05 y *** p < 0,01. *Fuente:* elaboración de los autores con información de la prueba Saber 3.º (Icfes), SIMAT (MEN) y registros del ICBF.

GRÁFICO 1: Desempeño en Saber 3.º por tipo de trayectoria de educación inicial



Nota: el gráfico presenta la diferencia estandarizada (*d* de Hedges) de cada grupo de trayectoria frente al grupo sin registro administrativo de educación inicial, que se ubica en cero por construcción. Valores positivos indican un desempeño superior al de ese grupo. Las diferencias se calculan incorporando valores plausibles, pesos muestrales y réplicas. ** p < 0,05 y *** p < 0,01. *Fuente:* elaboración de los autores con información de la prueba Saber 3.º (Icfes), SIMAT (MEN) y registros del ICBF.

Los promedios varían más entre tipos de prestador que entre el conjunto de estudiantes con y sin registro administrativo. La distancia entre el grupo con ICBF en establecimientos

oficiales y el grupo sin ICBF en establecimientos no oficiales se acerca a una desviación estándar. Esta diferencia es consistente con la selección socioeconómica de los estudiantes entre modalidades. En particular, los programas del ICBF priorizan a la población en mayor condición de vulnerabilidad. Por esta razón, la brecha observada no puede interpretarse como un efecto de la atención recibida.

3.2. Diferencias ajustadas por características observables

Con el fin de reducir las diferencias sistemáticas entre grupos, se construyeron comparaciones ajustadas mediante emparejamiento por puntaje de propensión, estimado por separado para los grupos con y sin participación en el ICBF. El modelo incluye variables que describen al estudiante y su contexto (autorreconocimiento cultural, región y municipio de nacimiento, estrato reportado durante la etapa preescolar) y los indicadores municipales de oferta descritos en la Subsección 2.3. El emparejamiento se realizó por vecino más cercano con reemplazo y una restricción de *caliper* de 0,2 desviaciones estándar del puntaje de propensión (Austin, 2011), siguiendo las recomendaciones de Caliendo y Kopeinig (2008). El número de vecinos se seleccionó por minimización del sesgo medio estandarizado y difiere entre grupos: un vecino en el grupo con ICBF y dos en el grupo sin ICBF. Los errores estándar incorporan simultáneamente el diseño muestral, la medición por valores plausibles y la dependencia inducida por la estructura de grupos emparejados. Esta última se ajusta por grupo emparejado y por identificador original del estudiante, pues el emparejamiento con reemplazo permite que una misma observación de control aparezca en varios grupos.

El ejercicio reduce sustancialmente el desbalance en las variables observables: el sesgo medio estandarizado pasa de 18,65 % a 3,01 % en el grupo con ICBF y de 7,49 % a 3,56 % en el grupo sin ICBF. Las diferencias ajustadas resultantes son, en general, considerablemente menores que las diferencias brutas de la Tabla 2. En el grupo con ICBF la diferencia en Lectura pasa de −28,69 a −6,50 puntos, y en el grupo sin ICBF pasa de +13,31 a +11,34 puntos. Al distinguir por naturaleza del establecimiento, la mayor diferencia positiva corresponde a los niños sin ICBF matriculados en establecimientos no oficiales, con 22,14 puntos en Lectura y 18,26 en Matemáticas.

Un ejercicio de sensibilidad adicional incorpora la incertidumbre asociada a la estimación del puntaje de propensión. Para ello, se reestima el modelo de participación y se repite el emparejamiento con cada réplica de pesos del diseño muestral. Con esta corrección, los errores estándar aumentan en todas las especificaciones. La diferencia del grupo con ICBF en Lectura deja de ser significativa a los niveles convencionales, mientras que las diferencias del grupo sin ICBF conservan su signo y significancia para las distintas cantidades de vecinos consideradas.

Estas magnitudes deben interpretarse con cautela y no corresponden a efectos causales. Primero, los modelos de participación tienen una capacidad discriminante moderada. El área bajo la curva ROC es 0,71 en el grupo con ICBF y 0,63 en el grupo sin ICBF. En

consecuencia, alcanzar un balance adecuado en las variables observadas no garantiza que se haya corregido la selección asociada a variables no observadas. Segundo, las diferencias pierden significancia estadística a los niveles convencionales cuando se incorporan covariables medidas después de la etapa preescolar. Este resultado también es consistente con la presencia de factores no observados relacionados tanto con la elección del tipo de atención como con el desempeño posterior.

4. Alcances y limitaciones de la base

Esta sección presenta las principales limitaciones de la base y precisa el alcance de las conclusiones que pueden derivarse de ella.

El grupo sin registro no es un contrafactual de ausencia de atención. Esta es la principal limitación del análisis. El grupo de 1.741 estudiantes se define por la ausencia de registro en SIMAT y en las bases del ICBF durante la etapa preescolar. La información disponible no permite establecer que estos estudiantes no hayan recibido educación inicial. Por ello, el grupo representa una categoría administrativa y no una condición de ausencia comprobada de atención. Su posición intermedia en la distribución de desempeño, documentada en la Sección 3, es compatible con la posibilidad de que algunos de estos niños hayan recibido atención en modalidades no observadas en las fuentes utilizadas. Entre estas modalidades podrían encontrarse jardines infantiles privados que no reportan al SIMAT, cuidado en el hogar con inversión parental o servicios prestados por cajas de compensación y programas territoriales que no aparecen en los registros consultados.

En consecuencia, las comparaciones que utilizan este grupo como referencia no representan el contraste entre recibir y no recibir educación inicial. Representan diferencias asociadas con la presencia o ausencia de un *registro administrativo* de atención, cuya interpretación es más limitada.

El tamaño y la composición del grupo de referencia restringen la precisión. El desbalance entre 1.741 estudiantes sin registro y 31.266 con registro obliga a reutilizar observaciones de control en el emparejamiento. En las especificaciones seleccionadas se utilizan 717 controles para el grupo con ICBF y 1.023 para el grupo sin ICBF. La observación con mayor reutilización sirve como comparación para 530 y 118 estudiantes tratados, respectivamente. La agrupación bidimensional de los errores estándar ajusta la inferencia por esta dependencia, pero no aumenta la información disponible para construir el grupo de comparación.

La restricción por edad excluye un resultado de interés. La base se restringe a estudiantes de ocho años o menos al momento de la aplicación, lo que excluye situaciones de extraedad y repitencia. La literatura internacional documenta que los efectos persistentes de la educación inicial pueden observarse con mayor claridad en indicadores de progresión

escolar, como la repitencia, la educación especial y la graduación, que en los puntajes de pruebas estandarizadas (McCoy et al., 2017; Gray-Lobe et al., 2023). Al restringir la muestra por edad, el ejercicio excluye una dimensión en la que la evidencia previa encuentra efectos. Además, si la educación inicial incide sobre la probabilidad de repetir, condicionar la muestra a quienes no repitieron introduce una selección sobre una variable posterior al tratamiento y afecta la comparabilidad de los grupos. Las versiones futuras de la base deberían conservar estas observaciones y analizar la extraedad y la repitencia como resultados.

Las dos cohortes no son directamente comparables. La etapa preescolar de los estudiantes evaluados en 2022 comprende los años 2016 a 2018, mientras que la de los evaluados en 2023 comprende los años 2018 a 2020. La emergencia sanitaria afectó a las dos cohortes en momentos diferentes. Para la primera, la suspensión de la presencialidad en marzo de 2020 ocurrió durante la educación básica primaria y afectó el aprendizaje posterior al preescolar. Para la segunda, 2020 corresponde al año teórico de transición, por lo que la interrupción afectó tanto la educación inicial como el registro de matrícula. Los datos son consistentes con esta diferencia. Entre 2022 y 2023, el promedio de Lectura disminuye 14,17 puntos en el grupo con registro administrativo y 5,83 puntos en el grupo sin registro. La primera diferencia es estadísticamente significativa y la segunda no. En consecuencia, incluir únicamente un control de año podría no recoger los cambios en la relación entre la trayectoria registrada y el desempeño. Los análisis futuros deberían estimar las dos cohortes por separado.

El tratamiento agrupa experiencias muy heterogéneas. La variable construida no distingue la duración, intensidad, continuidad o calidad de la atención. El grupo con participación en el ICBF reúne modalidades con propósitos y diseños pedagógicos diferentes, entre ellas los Hogares Comunitarios de Bienestar y los Centros de Desarrollo Infantil. La comparación experimental de estas modalidades ha encontrado efectos de distinto signo según la dimensión del desarrollo infantil analizada (Bernal et al., 2019). Por esta razón, su análisis agregado puede ocultar diferencias relevantes entre modalidades.

5. Potencial analítico y agenda de investigación

Las limitaciones anteriores delimitan las preguntas que pueden estudiarse con la base y permiten identificar cuatro líneas para ampliar su potencial analítico.

Extender los resultados hacia indicadores de trayectoria escolar. Desde 2017 no se realizan aplicaciones censales de las pruebas Saber en primaria, de modo que la medición de aprendizajes en ese nivel proviene hoy de muestras. Los indicadores de trayectoria de SIMAT, como el tránsito oportuno al grado de transición, la edad de ingreso, la extraedad, la repitencia, la aprobación, la deserción y la permanencia, son censales y anuales. Vincular la trayectoria de educación inicial con estos indicadores ampliaría la cobertura del análisis y

permitiría estudiar resultados para el universo de estudiantes registrados en SIMAT. También permitiría analizar dimensiones en las que la literatura internacional ha encontrado efectos persistentes.

Extender la ventana de cohortes. La restricción a las aplicaciones de 2022 y 2023 obedeció a la complejidad de estructurar los cruces y no a la disponibilidad de información. Extender el ejercicio hacia las aplicaciones censales de Saber 3.º y 5.º anteriores a 2017 permitiría observar cohortes cuya etapa preescolar coincide con la expansión de la política de atención integral a la primera infancia. La trayectoria de estas cohortes podría seguirse hasta Saber 11.º, una evaluación censal y anual. Ello abriría la posibilidad de contrastar resultados de corto y de largo plazo sobre las mismas cohortes.

Evaluar la variación en la oferta local como estrategia de identificación. La base ya incorpora indicadores municipales de cobertura de preescolar y de provisión de programas de primera infancia para el municipio y año de nacimiento de cada niño. Esa variación en la disponibilidad local de servicios se utiliza actualmente como covariable de ajuste. Estudios posteriores podrían evaluar si la expansión escalonada de la oferta pública genera una variación compatible con los supuestos requeridos para una estrategia de identificación. Este análisis permitiría estudiar qué grupos se benefician en mayor medida de la expansión de la oferta (Cornelissen et al., 2018).

Articular con los sistemas de seguimiento existentes. El Sistema de Seguimiento al Desarrollo Integral de la Primera Infancia, administrado por el Ministerio de Educación Nacional en articulación con la Comisión Intersectorial para la Primera Infancia, integra de manera nominal información de los registros del ICBF, del SIMAT y del sector salud. La articulación entre ese sistema y los registros de evaluación del Icfes permitiría seguir a las cohortes desde la atención en primera infancia hasta el final de la educación media, completando una trayectoria que hoy ninguna fuente individual observa.

Una consideración transversal a las cuatro líneas es la calidad del emparejamiento entre fuentes. El subregistro y las inconsistencias documentales pueden ser más frecuentes en zonas rurales dispersas y entre la población étnica o migrante. En consecuencia, las tasas de vinculación podrían variar entre subpoblaciones. Reportar estas tasas por zona, pertenencia étnica y condición socioeconómica permitiría evaluar la cobertura de la base e identificar posibles diferencias en la visibilidad administrativa de los niños colombianos.

6. Conclusiones

Esta nota documentó la construcción de una base que vincula los registros de matrícula del SIMAT, los registros de beneficiarios de programas de primera infancia del ICBF y los resultados de la prueba Saber 3.º de 2022 y 2023. La base reconstruye la trayectoria de educación inicial de 33.007 estudiantes, que representan a 821.721 niños. Hasta donde

conocemos, este es el primer ejercicio de esta naturaleza realizado con fines de investigación en Colombia. La información resultante puede sustentar otros análisis sobre primera infancia y trayectoria escolar.

Del ejercicio se desprenden tres conclusiones.

La primera se refiere a la cobertura de los registros administrativos. Los estudiantes sin registro administrativo de educación inicial no se ubican en el extremo inferior de la distribución de desempeño. Ocupan una posición intermedia y su promedio supera el de quienes registran atención en establecimientos oficiales. Por tanto, la ausencia de registro no puede interpretarse como ausencia de atención. Es posible que algunos niños reciban educación inicial en modalidades que los sistemas de información utilizados no observan. Cuantificar y caracterizar estas modalidades es una tarea pendiente para la planeación sectorial.

La segunda conclusión es que los promedios presentan una heterogeneidad importante según el tipo de prestador. La distancia entre los niños atendidos por programas del ICBF en establecimientos oficiales y los niños matriculados en establecimientos no oficiales sin participación en el ICBF se acerca a una desviación estándar. Esta brecha es consistente con diferencias socioeconómicas de origen y con los criterios de focalización de los programas del ICBF. Por esta razón, no puede atribuirse a la atención recibida.

La tercera conclusión es metodológica. La base permite caracterizar trayectorias de educación inicial con un nivel de detalle que no estaba disponible para Colombia, pero no permite estimar por sí sola el efecto causal de la educación inicial sobre el aprendizaje. El grupo de referencia se define por la ausencia de registro, existe un desbalance pronunciado entre grupos, la restricción por edad excluye la repitencia y las dos cohortes estuvieron expuestas a la emergencia sanitaria en momentos diferentes. Estos elementos no se resuelven únicamente mediante ajustes por variables observables. La base debe entenderse como una infraestructura para desarrollar nuevos análisis y diseños de identificación.

Referencias

Andrew, A., Attanasio, O., Fitzsimons, E., Grantham-McGregor, S., Meghir, C., y Rubio-Codina, M. (2018). Impacts 2 years after a scalable early childhood development intervention to increase psychosocial stimulation in the home: A follow-up of a cluster randomised controlled trial in Colombia. *PLOS Medicine*, 15(4), e1002556.

Andrew, A., Attanasio, O. P., Bernal, R., Cardona Sosa, L., Krutikova, S., y Rubio-Codina, M. (2024). Preschool quality and child development. *Journal of Political Economy*, 132(7).

Attanasio, O. P., Fernández, C., Fitzsimons, E. O. A., Grantham-McGregor, S. M., Meghir, C., y Rubio-Codina, M. (2014). Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: Cluster randomized controlled trial. *BMJ*, 349, g5785.

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, 10(2), 150–161.

Bernal, R. (2015). The impact of a vocational education program for childcare providers on children's well-being. *Economics of Education Review*, 48, 165–183.

Bernal, R., Fernández, C., Flórez, C. E., Gaviria, A., Ocampo, P. R., Samper, B., y Sánchez, F. (2009). *Evaluación de impacto del programa Hogares Comunitarios de Bienestar del ICBF*. Documentos CEDE, Universidad de los Andes.

Bernal, R., Attanasio, O., Peña, X., y Vera-Hernández, M. (2019). The effects of the transition from home-based childcare to childcare centers on children's health and development in Colombia. *Early Childhood Research Quarterly*, 47, 418–431.

Caliendo, M., y Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1), 31–72.

Cornelissen, T., Dustmann, C., Raute, A., y Schönberg, U. (2018). Who benefits from universal child care? Estimating marginal returns to early child care attendance. *Journal of Political Economy*, 126(6).

Evans, D. K., Jakiela, P., y Mendez Acosta, A. (2024). *The impacts of childcare interventions on children's outcomes in low- and middle-income countries: A systematic review*. CGD Working Paper 676. Center for Global Development.

Gray-Lobe, G., Pathak, P. A., y Walters, C. R. (2023). The long-term effects of universal preschool in Boston. *Quarterly Journal of Economics*, 138(1), 363–411.

McCoy, D. C., Yoshikawa, H., Ziol-Guest, K. M., Duncan, G. J., Schindler, H. S., Magnuson, K., Yang, R., Koepp, A., y Shonkoff, J. P. (2017). Impacts of early childhood education on medium- and long-term educational outcomes. *Educational Researcher*, 46(8), 474–487.