Análisis de la comparabilidad

Serie Aprender, nivel primario

Mayo 2024

Introducción

Los esfuerzos por evaluar los desempeños educativos en nuestro país se inician en la década de 1990 con el operativo nacional de evaluación (ONE) y continúan con las pruebas Aprender desde 2016. Las pruebas Aprender, en el nivel primario, evalúan la comprensión lectora (lengua) y la resolución de problemas o la solución de situaciones que resulten desafiantes (matemática), de la niñez escolarizada en Argentina. Los contenidos y capacidades que se evalúan en éstas se acuerdan y realizan en forma conjunta con los equipos técnicos de las jurisdicciones en procura de concordancia con los diseños curriculares de cada una.

Serie histórica de áreas evaluadas en primaria.

		2016	2017	2018	2021	2022	2023
Primaria	Censo	6° grado: Lengua y Matemática	6° grado: Ciencias Sociales y Ciencias Naturales	6° grado: Lengua y Matemática	6° grado: Lengua y Matemática		6° grado: Lengua y Matemática
	Muestra	3° grado: Lengua y Matemática	4° grado: Producción Escrita			6° grado: Lengua y Matemática	

Fuente: Evaluación Aprender 2023, DNEE-REFCEE | SIEE | Secretaría de Educación| Ministerio de Capital Humano

La última edición de Aprender se implementó en septiembre de 2023 a estudiantes de sexto grado de las escuelas primarias del país. En diciembre del mismo año se publicó una síntesis de resultados de la evaluación. Si bien no quedó registrado en el material publicado, esta

información se procesó con una base de datos muy avanzada pero no definitiva.

Durante diciembre, enero y febrero el equipo de metodología de la Dirección Nacional de Evaluación e Información Educativa siguió completando la base de datos y resolviendo problemas de la lectura óptica. La base final se cerró con 642.006 registros, 4.003 casos más que la usada para el informe preliminar.

Considerando el cambio de gestión, y que se estaba por entregar a los ministros de cada jurisdicción los resultados de un dispositivo que no se había dirigido por esta gestión, se decidió solicitar a un consultor externo la revisión de la calidad del trabajo realizado. En esta revisión se contempló:

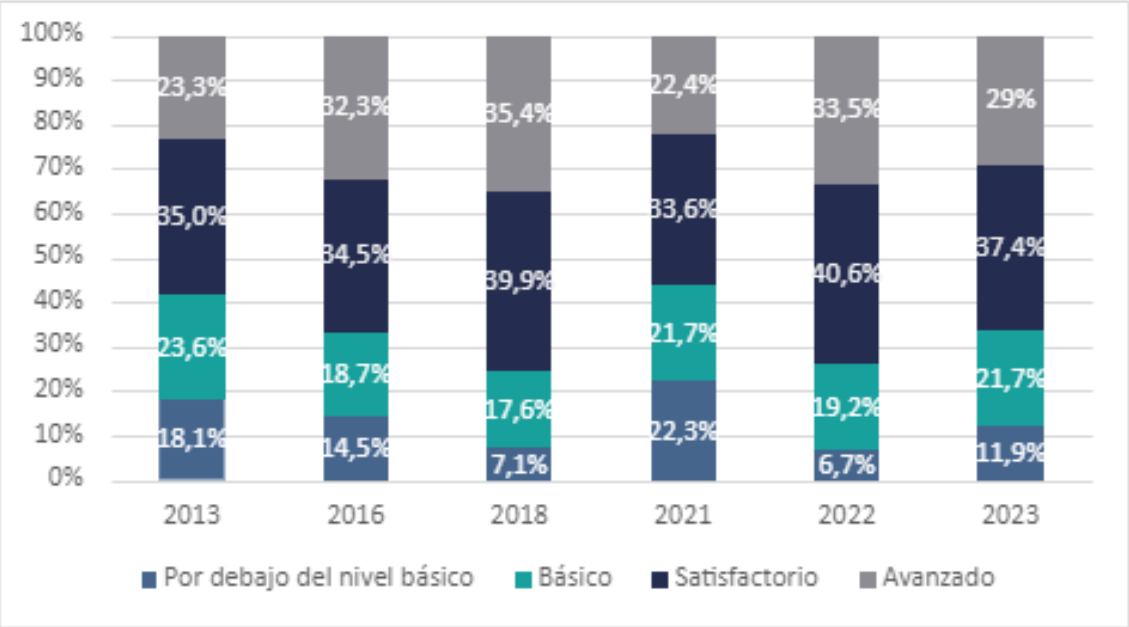
- En cuanto a los aspectos cognitivos se revisó el funcionamiento psicométrico de los ítems (discriminación y dificultad), si había diferencias notorias de dificultad entre los diferentes modelos, y el cálculo del puntaje a través de TRI. Se concluyó que no había habido problemas en la calidad del trabajo.
- En el caso de los cuestionarios complementarios se decidió realizar una imputación de la no respuesta en las variables vinculadas a la construcción del nivel socioeconómico, dada la alta no respuesta registrada en algunas de estas variables.
- Realizada esta imputación en marzo se enviaron las bases de datos definitivas y un documento técnico a los ministros jurisdiccionales.

En segunda instancia se comenzó a elaborar el Informe Aprender. En este informe se presenta información detallada de los desempeños y factores asociados en el 2023, a la vez que se presenta una evolución de los niveles de desempeño en cada área evaluada a lo largo del tiempo.

La serie construida desde 2013 evidencia una tendencia ascendente en los niveles de competencia en lengua entre 2013 y 2018, dada la disminución progresiva del porcentaje de estudiantes que no alcanzan el nivel satisfactorio. Durante la pandemia, se registra un deterioro notable en el rendimiento: el porcentaje de alumnos que no alcanzan el nivel satisfactorio aumenta del 25% en 2018 al 45% en 2021. Después, en un año, se observa una mejora importante, reduciendo este porcentaje

al 25,9 % en 2022. En el año 2023 el nivel de los aprendizajes vuelve descender.

Gráfico 1. Evolución histórica del nivel de desempeño en Lengua en 6° grado de Nivel Primario. ONE 2013 y Aprender 2016-2023.



Fuente: Evaluaciones ONE 2013, Aprender 2016, 2018, 2021, 2022 y 2023, DEE-REFCEE | DiNEIEE | SEIE | Ministerio de Educación de la Nación.

La dificultad para interpretar la magnitud de estos avances y retrocesos en el desempeño educativo generó la necesidad de evaluar si estas variaciones eran reales en la población, o si pudieran ser atribuibles a algún error de medición. La Subsecretaría de Información y Evaluación Educativa, a través de la Dirección Nacional de Planeamiento e Investigación Educativa y la Dirección Nacional De Evaluación e Información Educativa convocó a dos especialistas para que analicen la serie de resultados en lengua y elaboren un informe. Estos consultores contaron las distintas bases de datos y con acceso a los equipos técnicos para consultar todos los protocolos de trabajo que consideraran necesarios.

Se trabajó sobre distintas hipótesis de errores de medición que podrían explicar la magnitud de las diferencias en el tramo 2018-2021 y 2021-2023 en la serie del porcentaje de alumnos de sexto grado lengua por debajo del básico.

- El diseño y selección de la muestra 2022, afectando a la validez externa del diseño.
- Problemas de precisión de los instrumentos, en concreto, los índices de dificultad y de discriminación de los ítems.
- Diferencias en las dificultades de los modelos.
- Problemas en el cálculo de los puntajes de cada alumno y parámetros de los ítems, con los ítems finales.

Este documento sintetiza los hallazgos de los consultores. En primer lugar, se describen de manera sucinta las características de las pruebas de lengua en el nivel primario. En segundo término, se analizan los diferentes aspectos que podrían haber incidido en las variaciones de la serie. Por último, se presentan conclusiones y recomendaciones así como la serie corregida.

Informe técnico de serie histórica Aprender

Augusto Hoszowski

María Elena Brenla

1. Características de las pruebas Aprender de Lengua en el nivel primario

En la Evaluación Nacional Aprender de Lengua en el nivel primario se evalúa la comprensión lectora de estudiantes de 6° grado a partir de la lectura de cuentos de literatura infantil, biografías, artículos periodísticos de interés general y artículos expositivos extraídos de enciclopedias infantiles, de breve extensión.

Para dar cuenta de cómo los alumnos comprenden estos tipos de textos se elaboraron preguntas que miden su desempeño lector en tres capacidades: extraer información explícita, interpretar a partir de inferencias significados locales y globales de los textos, y reflexionar y evaluar sobre el contenido y la forma de los textos desde sus conocimientos previos. Los contenidos escolares evaluados a partir de estas tres capacidades son, entre otros: información literal, macroestructura textual, género discursivo, trama textual, paratextos, idea central, especificidad del texto literario, vocabulario, recursos enunciativos y elementos de cohesión.

De cada texto evaluado se desprenden doce ítems de opción múltiple con cuatro opciones de respuesta y una sola respuesta correcta.

Cada estudiante responde un cuadernillo de prueba con un total de dos textos y veinticuatro preguntas.

Los ítems tienen diferentes grados de dificultad, lo que permite establecer cuatro niveles de desempeño diferenciado (Nivel por debajo del Básico, Básico, Satisfactorio y Avanzado), de acuerdo con cómo han sido las respuestas de los y las estudiantes y sobre la base de los criterios establecidos mediante la metodología bookmark, realizado con referentes docentes de todo el país en 2016.

El formato de estos cuestionarios se mantiene a lo largo de los distintos operativos.

Los cuestionarios Aprender siguen en general el siguiente esquema:

Con seis bloques de 12 ítems cada uno, se construyen seis modelos de 24 ítems cada uno:

	Modelos					
	1	2	3	4	5	6
Bloques	1	2	3	4	5	6
	4	5	6	2	3	1

Por ejemplo, el modelo 5 tiene en primer lugar (los primeros doce ítems) al bloque 5 y en segundo lugar al bloque 3

- Cada bloque aparece en dos modelos: en uno en primer lugar y en otro en segundo lugar. Para evitar así un 'efecto posición'
- No todos los pares de bloques aparecen en un mismo modelo. Por ejemplo, los bloques 1 y 2 no aparecen juntos en ningún modelo

Un mínimo de dos de los seis bloques, en modelos distintos, se repiten a lo largo de los operativos Aprender para permitir la comparación de los puntajes.

1.1. Equiparación de las puntuaciones: anclajes

Dado que las pruebas Aprender se aplican en forma periódica es importante definir el método de equiparación de sus puntuaciones a lo largo del tiempo. Uno de los diseños más usados en las pruebas educativas a gran escala es el de grupo no equivalentes con test de anclaje. Esta metodología es la que se ha adoptado en Argentina desde las pruebas ONE a las Aprender. Consiste en administrar, en cada evaluación, un grupo de ítems llamado "test de anclaje" que permite establecer la equivalencia de las pruebas en forma longitudinal. En lo habitual, se suele incluir al menos un 20% de ítems comunes en el total de la prueba. En el caso de Lengua, cada anclaje está compuesto por un bloque de 12 ítems que refieren a un mismo texto.

Aprender 2021 contó con dos bloques de anclaje: un anclaje histórico que viene de las pruebas ONE y un anclaje Aprender. Aprender 2022 sumó un tercer anclaje que se aplicó por vez primera en 2021. Aprender 2023 contó con 3 anclajes.

1.2. Teoría de Respuesta a Ítem

Las pruebas Aprender, como muchas de las evaluaciones educativas estandarizadas, se procesan mediante la Teoría de Respuesta al Ítem (TRI).

A diferencia de los procesamientos tradicionales (Teoría Clásica), que consisten en sumar o contabilizar las respuestas correctas (en el caso de preguntas con respuestas codificadas V/F), el modelo TRI supone que cada alumno posee cierta competencia representada por un número real (θ) y que, en el modelo aplicado por Aprender, cada ítem está caracterizado por su *dificultad* y su *discriminación*. Tanto los puntajes de los alumnos como los parámetros de los ítems se pueden ubicar en la misma escala, que es en cierto sentido arbitraria. En Aprender 2016 la media de la escala se fijó en 500 y el desvío estándar en 100.

El procesamiento de una encuesta Aprender con TRI produce un puntaje para cada alumno (Theta) y los parámetros Dificultad y Discriminación para cada ítem. Este valor Theta no dice si el alumno alcanzó o no un nivel de desempeño aceptable o superior al mínimo deseado. Para ello hay que interpretar la escala de puntajes, estableciendo *puntos de corte* que definan niveles de desempeño. Esto se hizo en el operativo Bookmark en Aprender 2016, que definió tres puntos de corte, para delimitar cuatro niveles de desempeño:

Por debajo del básico – Básico – Satisfactorio - Avanzado

Esto puntos de corte se mantienen fijos en la escala definida para permitir la comparación de resultados a lo largo del tiempo.

1.3. Dificultades de los modelos

La construcción de los modelos considera sus contenidos y su dificultad. Al construir los modelos se toma en cuenta los resultados de las pruebas piloto, en donde se prueban los ítems para estimar sus

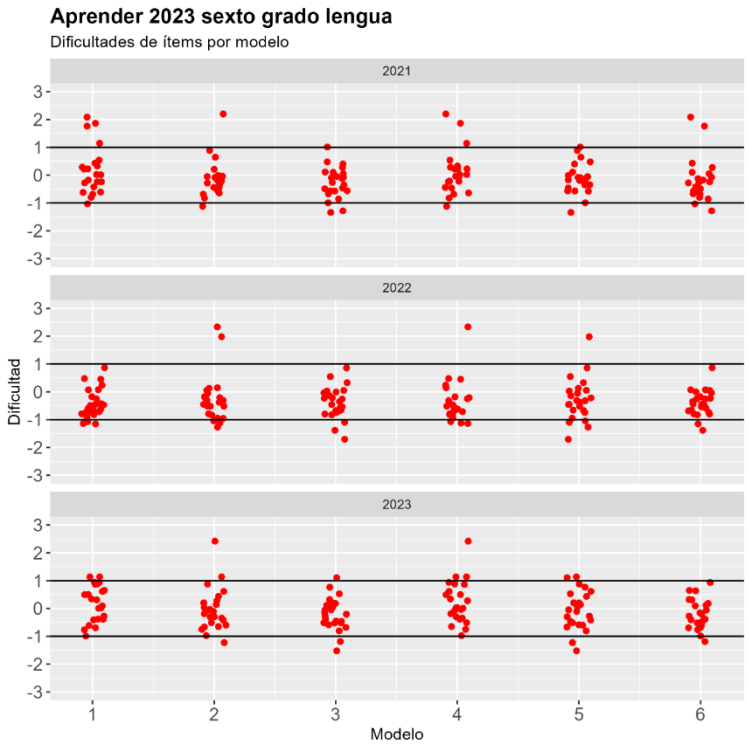
niveles de dificultad y discriminación. Aquellos ítems que no pasan las pruebas psicométricas son desechados en esta etapa.

2. Análisis de la calidad de las pruebas

En primer lugar, se analizó la muestra Aprender 2022. El operativo 2022 tiene como tamaño de muestra efectivo en lengua 3.682 colegios a nivel nacional, muy por encima de lo mínimo recomendado para este tipo de operativos. Y más de 80 colegios por jurisdicción (479 en Buenos Aires y 143 en CABA), garantizando estimaciones confiables a nivel nacional. Las características de la muestra 2022 están bien documentadas y no se encontró ningún procedimiento objetable.

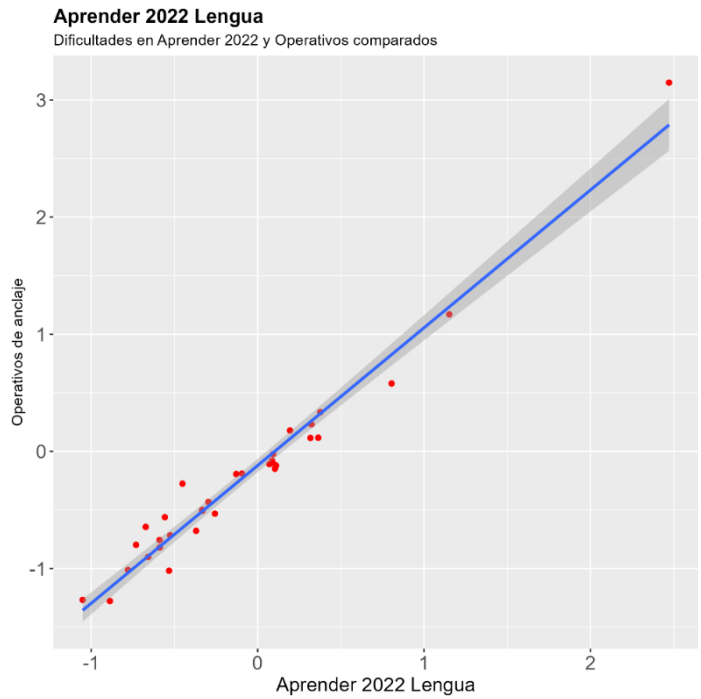
En segundo lugar, se analizó el funcionamiento psicométrico y diferencias en la dificultad entre modelos. Tampoco se encontró nada para objetar.

Como ejemplo, se grafican las dificultades por modelo de los ítems de Aprender 2021, 2022 y 2023 lengua sexto grado.



Se observa que a lo largo de los operativos la distribución de dificultades de los modelos es similar. El puntaje medio en cada operativo se estima con los seis modelos en conjunto.

Se revisó si hubo ítems en el bloque de anclaje que de un operativo a otro hayan variado sus parámetros en forma sospechosa. No se encontró nada objetable.



Una vez controlado las variaciones de dificultad y discriminación, se controlaron los puntajes escalados.

Cálculo de la proporción de alumnos en cada nivel de desempeño.

Escalamiento de las encuestas Aprender

Una vez procesada una prueba para comparar los puntajes de dos operativos (de una misma disciplina y grado escolar) y poder evaluar si las competencias de los alumnos aumentaron o disminuyeron, es necesario llevar los puntajes a una escala prefijada. En nuestro caso la de Aprender 2016: media 500 y desvío estándar 100. El método utilizado

por Aprender se denomina **mean-mean** (toma en cuenta la media de las dificultades y la media de las discriminaciones en el bloque de anclaje).

Los puntajes de cada operativo se llevan luego a la escala 2016 en base a la comparación de las respuestas de los bloques de anclaje:

2.1. Respuestas a las mismas preguntas

Como resultado, la media del puntaje de cada operativo aumenta o disminuye según cómo el bloque de anclaje se haya comportado.

La media de puntaje escalado de un operativo no está influenciada por los ítems que no sean comunes.

En todas las encuestas similares a Aprender el escalamiento de un año a otro (para una misma disciplina y grado escolar) toma en cuenta solamente los ítems de anclaje.

2.2. Puntaje medio y alumnos en los niveles de desempeño

Algo que también se observa es que la proporción de alumnos en los niveles de desempeño 'Por debajo del básico' y 'Por debajo del básico' + 'Básico' están correlacionados con el puntaje medio:

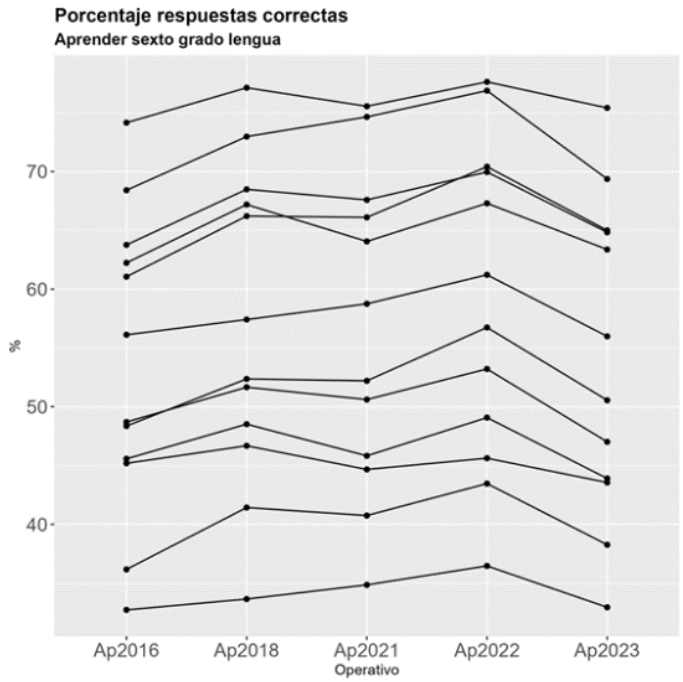
- A mayor puntaje medio, menor proporción de alumnos que no alcanzan el nivel Satisfactorio
- A menor puntaje medio, mayor proporción de alumnos que no alcanzan el nivel Satisfactorio

2.3. Puntaje TRI y proporción de respuestas correctas

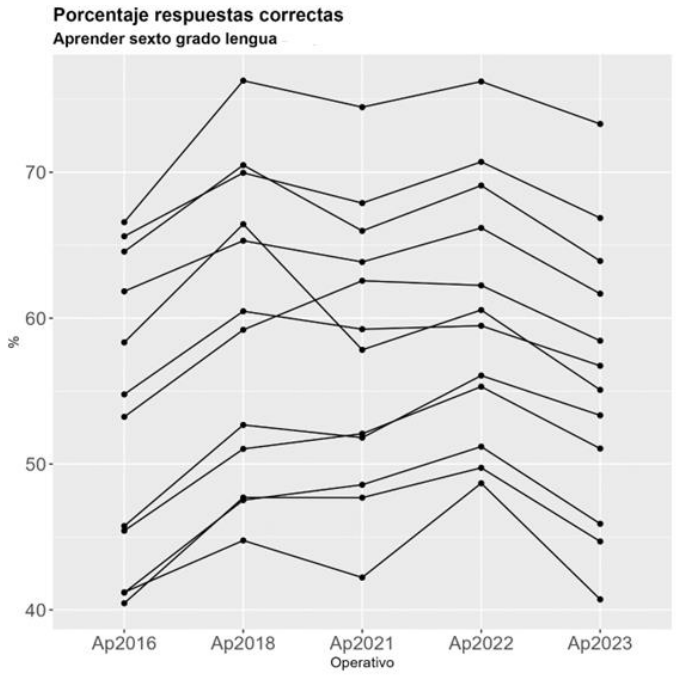
De un operativo a otro, el aumento o disminución del puntaje medio (*escalado*) está aproximadamente en correspondencia con la cantidad de ítems en común correctamente respondidos.

Los siguientes gráficos muestra el porcentaje de respuestas correctas a dos conjuntos de ítems en común, en lengua sexto grado.

Conjunto bloque de anclaje A



Conjunto bloque de anclaje B



Recordamos que son los mismos ítems aplicados en los operativos 2016, 2018, 2021, 2022, 2023. Y ubicados en la misma posición en los

cuestionarios. La fecha (en cada año) de aplicación en los cinco operativos es similar. No se incluyó el tercer bloque para poder graficar la serie desde 2016.

Con diferencias de niveles la forma de las líneas se mantiene entre ítems y entre ambos bloques:

- Disminución de 2018 a 2021
- Aumento de 2021 a 2022
- Disminución de 2022 a 2023

2.4. Escalamiento de la prueba de sexto grado 2021

Los gráficos anteriores, con la proporción de respuestas correctas en dos bloques de anclaje, permitieron detectar un error en 2021, ya que los puntajes y niveles de desempeño publicados eran incoherentes con las respuestas correctas en los bloques de anclaje.

El origen del error fue que en 2021 se invirtió el orden de los procesos de escalamiento y equiparación. Luego de esto, se omitió agregar un factor de corrección en el momento de la equiparación, ya que esta se hizo con media 0 y no con media 500. Este paso que no se usó impactó en la varianza y en el puntaje final.

Al detectarse este error se recalculó la serie de puntajes y niveles de desempeño.

2.5. Series de niveles de desempeño y valor erróneo 2021

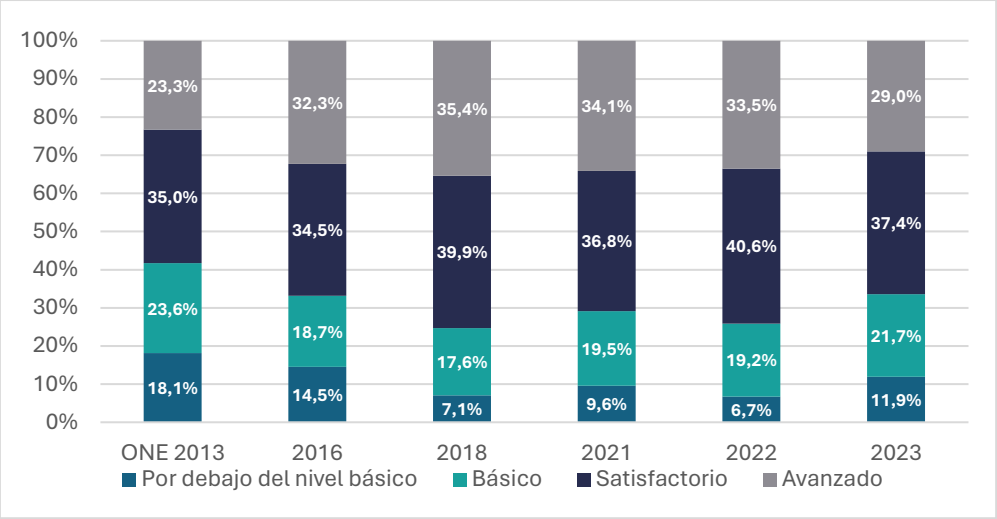
En las siguientes series se incluye para el año 2021 sexto grado lengua y sexto grado matemática tanto el valor correcto como el erróneamente informado. Se incluye el año 2013, que corresponde al Operativo ONE, a título informativo. Pero por diversos motivos (un solo bloque de anclaje entre este operativo y Aprender, ausencia de escuelas privadas en Córdoba en el operativo sexto grado, desbalanceo entre sectores de gestión en CABA, etc.) la comparación debe ser hecha con precaución.

3. Conclusiones y recomendaciones

- A lo largo de los operativos todos los ítems que se retuvieron para calcular los puntajes funcionaron correctamente desde el punto de vista psicométrico.
- La cantidad de ítems utilizados para el escalamiento está en el orden sugerido por los expertos.
- La distribución de dificultades de los ítems es similar en todos los operativos.
- El Operativo 2022 tiene como tamaño de muestra efectivo en lengua 3682 colegios a nivel nacional, muy por encima de lo mínimo recomendado para este tipo de operativos. Y más de 80 colegios por jurisdicción (479 en Buenos Aires y 143 en CABA), garantizando estimaciones confiables a nivel nacional.
- Un tercer bloque de anclaje aparece en lengua sexto grado 2022 (se aplica por primera vez en 2021). Este bloque presenta menores dificultades que los otros. Esto, en el modelo TRI, no influye en el escalamiento. Además, no es que se suplantó un bloque por otro, sino que es un bloque que se adicionó como bloque en común, mejorando con ello la comparabilidad.
- En el Operativo 2021 los seis modelos estiman los niveles de desempeño en forma similar al resto.
- La distribución de los puntajes en Aprender 2021 Lengua es similar en los tres primeros modelos.
- En 2021 hubo un problema técnico de escalamiento que impactó en el cálculo de los puntajes y del porcentaje de alumnos de cada nivel de desempeño. Este error debe corregirse para poder recalcular la serie.
- Se deben corregir los resultados de acceso público derivados de este error.
- Se deben incorporar más protocolos de chequeo de la información en diferentes momentos del procesamiento de las pruebas.

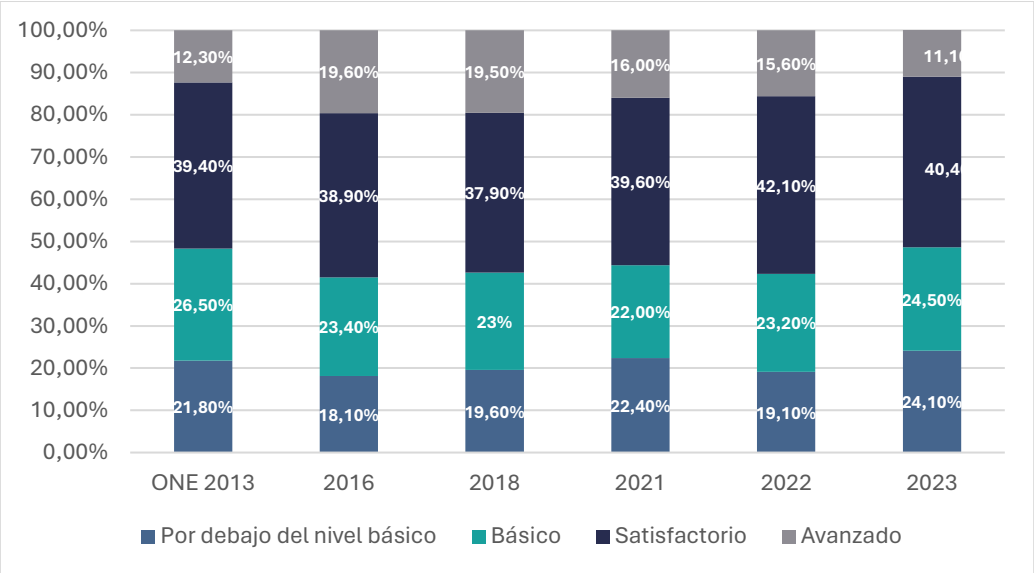
4. Series corregidas

Evolución del nivel de desempeño en Lengua en Aprender 6° grado de nivel primario



Fuente: Evaluación Aprender 2023, DNEE-REFCEE | SIEE | Secretaría de Educación| Ministerio de Capital Humano

Evolución del nivel de desempeño en Matemática en Aprender 6° grado nivel primario



Fuente: Evaluación Aprender 2023, DNEE-REFCEE | SIEE | Secretaría de Educación| Ministerio de Capital Humano